UNIVERSITY OF CALIFORNIA

Los Angeles

Qualitative Probabilities:
A Normative Framework for Commonsense Reasoning

A dissertation submitted in partial satisfaction
of the requirements for the degree
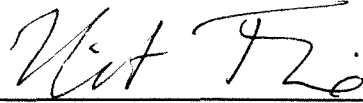Doctor of Philosophy in Computer Science
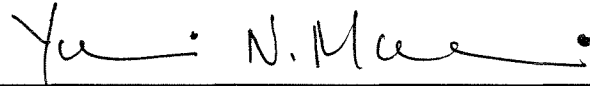
by

Moisés Goldszmidt

1992

Report Documentation Page

1. REPORT DATE **1992**	2. REPORT TYPE	3. DATES COVERED **00-00-1992 to 00-00-1992**

4. TITLE AND SUBTITLE **Qualitative Probabilities: A Normative Framework for Commonsense Reasoning**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California Los Angeles,Department of Computer Science,Los Angeles,CA,90095**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Intelligent agents are expected to generate plausible predictions and explanations in partially unknown and highly dynamic environments. Thus, they should be able to retract old conclusions in light of new evidence and to efficiently manage wide fluctuations of uncertainty. Neither mathematical logic nor numerical probability fully accommodates these requirements. In this dissertation I propose a formalism that facilitates reasoning with qualitative rules, facts, and deductively closed beliefs (as in logic), yet permits us to retract beliefs in response to changing contexts and imprecise observations (as in probability). Domain knowledge is encoded as if-then rules admitting exceptions with different degrees of abnormality, and queries specify contexts with different levels of precision. I develop effective procedures for testing the consistency of such knowledge bases and for computing whether (and to what degree) a given query is confirmed or denied. These procedures require a polynomial number of propositional satisfiability tests and hence are tractable for Horn expressions. Finally, I show how to give rules causal character by enforcing a Markovian condition of independence. The resulting formalism provides the necessary machinery for embodying belief updates and belief revision, generating explanations, and reasoning about actions and change.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a REPORT **unclassified**	b ABSTRACT **unclassified**	c THIS PAGE **unclassified**	**Same as Report (SAR)**	**174**	
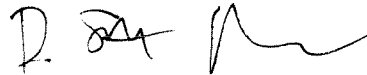
The dissertation of Moisés Goldszmidt is approved.

_____
Kit Fine
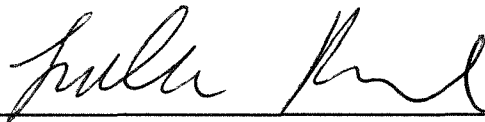
_____
Yiannis Moschovakis

_____
D. Stott Parker

_____
Sheila A. Greibach

_____
Judea Pearl, Committee Chair

University of California, Los Angeles

1992

TABLE OF CONTENTS

LIST OF FIGURES

List of Tables

ACKNOWLEDGMENTS

my grandparents, Dora and Paul, Enja and Isaac; my brothers and sisters, Freddy, Helen, Alexandra (in Italy), Leon, William and Ricardo. My deepest gratitude goes to my parents, Ada and Albert, for believing in me, always being there, making things possible, and constantly proving that nothing comes before the well-being of their children.

Almost every memory of happiness from my childhood is laced with the musical voice and laughter of my grandmother, Dora, who passed away during the summer of 1990. She is the only reason that would make me want to believe in heaven. Abuela Dora, I miss you.

Tania – best friend, compinche, and my beautiful wife – has the miraculous ability to turn worry into fun. She has encouraged and participated in every one of my (big and small) projects and has been an uncompromising critic of my work (and my music). Tania I love you. Thanks!

The music is by Charles Mingus.

VITA

1960	Born, Caracas, Venezuela.
1983	Electronic Engineer – Cum Laude –, Universidad Simon Bolivar, Caracas, Venezuela
1983-1985	Research and Development Engineer, Venezuelan Institute of Scientific Research (I.V.I.C. – F.I.I.)
1985–1986	Teaching Assistant, Electrical Engineering Department, UCSB.
1986–1987	Research Assistant, Computer Science Department, UCSB.
1987	M.S. Electrical Engineering – Computer Science Track, UCSB, Santa Barbara, California
1987–1992	Research Assistant, Computer Science Department, Cognitive Systems Laboratory, UCLA.
1990–1992	Recipient of an IBM Graduate Fellowship.
1992	Ph.D. Computer Science, UCLA, Los Angeles, California

PUBLICATIONS

"A Maximum Entropy Approach to Nonmonotonic Reasoning", Moisés Goldszmidt, Paul Morris, and Judea Pearl, *IEEE Pattern Analysis and Machine Intelligence*, (in press) 1992. A short version can be found in *Proceedings of American Association for Artificial Intelligence Conference*, pages 646–652, Boston, 1990.

"Rank-Based Systems: A Simple Approach to Belief Revision, Belief Update, and Reasoning About Evidence and Actions", Moisés Goldszmidt and Judea Pearl, to appear in Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning, KR-1992.

"Reasoning with Qualitative Probabilities Can Be Tractable", Moisés Goldszmidt and Judea Pearl, in *Proceedings of the 8th Conference on Uncertainty in Artificial, Intelligence*, pages 112–120, Stanford, CA, 1992.

"Stratified Rankings for Causal Modeling", Moisés Goldszmidt and Judea Pearl, in *Proceedings of the Fourth International Workshop on Nonmonotonic Reasoning*, pages 99–110, Vermont, 1992.

"On The Consistency of Defeasible Databases", Moisés Goldszmidt and Judea Pearl, *Artificial Intelligence*, Vol. 52:2, pages 121–149, 1991.

"System Z^+: A Formalism for Reasoning with Variable Strength Defaults", Moisés Goldszmidt and Judea Pearl, in *Proceedings of American Association for Artificial Intelligence Conference*, pages 399–404, Anaheim, CA, 1991.

"On The Relation Between Rational Closure and System-Z", Moisés Goldszmidt and Judea Pearl, in *Third International Workshop on Nonmonotonic Reasoning*, pages 130–140, South Lake Tahoe, CA, 1990.

"Deciding Consistency of Databases Containing Defeasible and Strict Information", Moisés Goldszmidt and Judea Pearl, in M. Henrion et. al., editor, *Uncertainty in Artificial Intelligence (Vol. 5)*. North Holland, Amsterdam, 1990. Also in the *UCLA Annual Research Review* 1990.

Qualitative Probabilities:
A Normative Framework for Commonsense Reasoning

by

Moisés Goldszmidt

Doctor of Philosophy in Computer Science
University of California, Los Angeles, 1992
Professor Judea Pearl, Chair

Intelligent agents are expected to generate plausible predictions and explanations in partially unknown and highly dynamic environments. Thus, they should be able to retract old conclusions in light of new evidence and to efficiently manage wide fluctuations of uncertainty. Neither mathematical logic nor numerical probability fully accommodates these requirements.

In this dissertation I propose a formalism that facilitates reasoning with qualitative rules, facts, and deductively closed beliefs (as in logic), yet permits us to retract beliefs in response to changing contexts and imprecise observations (as in probability). Domain knowledge is encoded as if-then rules admitting exceptions with different degrees of abnormality, and queries specify contexts with different levels of precision. I develop effective procedures for testing the consistency of such knowledge bases and for computing whether (and to what degree) a given query is confirmed or denied. These procedures require a polynomial number of propositional satisfiability tests and hence are tractable for Horn expressions. Finally, I show how to give rules causal character by enforcing a Markovian condition of independence. The resulting formalism provides the necessary machinery for embodying belief updates and belief revision, generating explanations, and reasoning about actions and change.

CHAPTER 1

Introduction

In their everyday interactions with the world, people continuously jump to conclusions on the basis of imperfect and defeasible information. For example, we normally expect to find our car where we parked it last, and upon turning the ignition key, we expect the engine to start. These expectations are plausible but not provable from what is known at the time they are assessed, and they may be replaced as new evidence is encountered. A stolen car will not be where we parked it last. An engine with a dead battery will not start. Yet, despite the multitude of possible scenarios, people operate under a fairly uniform *consensus* as to what is plausible, that is, what should be upheld as true for practical purposes. This suggests that there are simple principles that govern the dynamics of plausible reasoning, including the distinction between plausible and implausible conclusions.

This dissertation is concerned with casting the principles governing plausible reasoning in a formal language. We wish to create through such formalization programs capable both of accepting and organizing input ranging from defeasible information such as "typically, if we turn the car's ignition the engine starts" to nondefeasible (strict) information such as "all humans are mortal" and of answering queries about what would be a plausible conclusion given some particular context. Figure 1.1 presents a schematic of this project. The set of "if φ_i then ψ_i" rules, $\varphi_i \xrightarrow{\delta_i} \psi_i$, represents a knowledge base encoding information about the world. The incompleteness of this information is modeled by allowing exceptions to these rules, where δ_i represents the *degree of abnormality* of these exceptions. Rules expressing what is normally the case without excluding the possibility of exceptions, are commonly known in Artificial Intelligence (AI) as *default rules*.[1] A query is a pair (ϕ, σ) representing the context ϕ and the target σ. The context of a query contains factual information on what is currently known about the environment, which may originate from either passive observations or active manipulations.[2] The target σ is a propositional hypothesis representing the

---

[1] In the database literature, these rules play the role of integrity constraints, but are normally treated as hard laws, tolerating no exceptions [111].

[2] This distinction is of crucial importance, as shown in Section 5.4.

1

Knowledge Base:

$$\Delta = \begin{array}{l} \varphi_1 \xrightarrow{\delta_1} \psi_1 \\ \varphi_2 \xrightarrow{\delta_2} \psi_2 \\ \vdots \\ \varphi_n \xrightarrow{\delta_n} \psi_n \end{array}$$

Confirmed

σ (degree?)

Denied

ϕ: Actions – Observations   Query: (ϕ, σ)
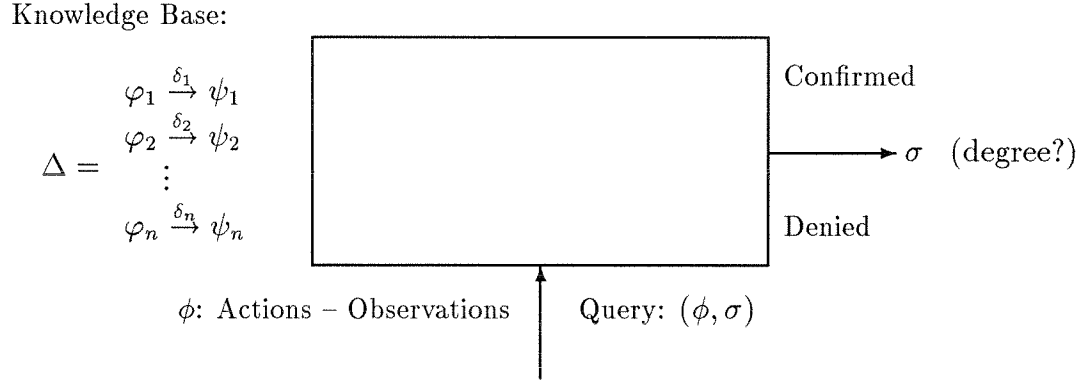
Figure 1.1: Schematic of the system proposed.

agent's current interest. The output is a decision about whether σ is plausible (and to what degree) given that ϕ is true.

Recently, defaults have been proposed in AI for expressing commonsense knowledge [109, 45]. Inheritance hierarchies, for instance, encode the prototypical properties of classes as defaults. In reasoning about change, defaults encode the tendency of properties to remain invariant in the absence of relevant changes. In diagnostic reasoning, defaults encode the rarity of faulty components. Even deductive databases usually embed default "closed-world" assumptions to fill in missing information. In short, any activity for which we cannot afford to specify in advance all responses to all conceivable situations seems to require use of defaults.

Initial attempts in AI to formalize plausible reasoning based on default rules favored extensions of classical logic [108, 87, 88, 90] to account for the *nonmonotonicity* of the default-based inferences.[3] Deduction in classical logic is monotonic: given that C is entailed by a theory T, C is also entailed by a theory T', where T' is a superset of T. On the other hand, inferences based on default rules are nonmonotonic: For example, I *believe/infer* that my car's engine will start (C) once I turn the ignition key (T), but would like to retract this belief (and infer that it will not start, i.e., $\neg C$) if the battery is dead (T'). Although these extensions successfully reproduced this nonmonotonic behavior, the interactions among defaults rules yield conflicting and sometimes couterintuitive conclusions.

---

[3]These formalisms are reviewed in Section 1.2.

2

For example, consider a knowledge base containing the defaults: "typically penguins don't fly", "typically birds fly", and the nondefeasible rule "all penguins are birds". Given that Tweety is a penguin we may conclude that Tweety does not fly based on the information provided by the first default. On the other hand, since Tweety is a penguin, she is also a bird, we may conclude that Tweety flies using the second default rule. The reason we prefer to uphold the conclusion that Tweety does not fly is based on the intuition that defaults providing information about a more specific class of individuals (i.e., penguins in this case) should be considered with a higher priority. There are other cases of default interactions and criteria for avoiding undesirable inferences based on assumptions of *minimal change*, and on notions of causality and explanation [57, 36, 8, 45, 121, 54, 118]. Some proposals address the problem of default interactions by asking the user to explicitly specify preferences among rules (e.g., [112, 30, 88, 80]). Ideally, however, such information should be extracted from the rules themselves (or their semantical interpretation), since anticipating interactions among defaults becomes increasingly difficult as the size of the knowledge base grows. Furthermore, user specification of these preferences seems to require a possibly exhaustive enumeration of cases, situations, and exceptions, which was precisely what the use of default rules meant to avoid. A related problem is the observation that some plausible conclusions are *harder* to retract than others in the face of conflicting evidence. This observation suggests that semantical interpretations of plausible beliefs should involve *rankings* or *orderings* among these beliefs (in addition to truth values) [33, 34, 120].

On the positive side a formalization in terms of a logical framework offers several advantages such as: independence of a specific implementation or domain and the possibilities of model theoretic interpretations and well-founded semantics.

On the other extreme, an alternative to extending classical logic may be probability theory: Uncertainty can be used to represent both the incompleteness of the information in the knowledge base and numbers can be used to model rankings of beliefs. Furthermore, Bayesian conditioning offers a successful and well understood method of dealing with retractions and belief change. Yet, a straightforward probabilistic interpretation of plausibility in terms of numbers and thresholds will encounter obstacles of its own. First, the picture we form about our environment seems to be encoded in terms of *plain beliefs*, that is, propositions that are accepted as true (for practical purposes), and continue to guide our actions until refuted by new evidence. These propositions are transmitted linguistically, and are qualified by expressions such as such as "generally", "extremely typical", and "very likely", which are void of precise numerical value.

Second, plain beliefs also seem to be deductively closed: If ψ is believed and φ is believed, then $\psi \wedge \varphi$ is believed as well. Note that if we associate the acceptance of ψ as a believed proposition with $P(\psi) > t$, where t is some suitable threshold, it is possible to have both $P(\psi) > t$ and $P(\varphi) > t$ but $P(\psi \wedge \varphi) < t$.

In this thesis, I propose a *conditional interpretation* of the default rules that presents the merits of both logic and probability. The sentence "if φ then ψ" is interpreted as imposing a preference for accepting ψ over $\neg\psi$ if φ is all that is known. This interpretation is based on an abstraction of probability theory where "if φ then ψ" constrains the conditional probability of ψ given φ to be infinitesimally close to 1. Intuitively, this amounts to according the consequence ψ a very high likelihood when φ is all we know.[4] As will be seen, at the heart of this formulation is the concept of *default priorities*, namely, a natural ordering of the rules which is derived automatically from the knowledge base. Representing and reasoning with causal relations is enabled through a stratified set of (probabilistic) independencies based on Markovian considerations. The result is a model-theoretic and semantically well-founded account of plausible beliefs that, as in classical logic, are qualitative and deductively closed and, as in probability, are subject to retraction and to varying degrees of firmness.

1.1   Overview and Summary of Contributions

Attaching probabilistic semantics to conditional sentences (i.e., if–then expressions) goes back to Adams [1, 2], who developed a logic of indicative conditionals based on infinitesimal probabilities. This logic includes a norm of consistency, called *p-consistency*, which tolerates exceptions (e.g., "typically, if φ then ψ and if $\varphi \wedge \varphi'$ then $\neg\psi$") and rules out contradictions (e.g., the pair "typically, if φ then ψ" and "typically, if φ then $\neg\psi$"). It also admits a notion of entailment, called *p-entailment*, which guarantees arbitrarily high probabilities for the conclusions whenever sufficiently high probabilities can be consistently assigned to the premises.

Unfortunately, Adams' notions of p-consistency and p-entailment were restricted to knowledge bases containing only defeasible information. The first contribution of this thesis is the extension of Adams' consistency and entailment to handle both defeasible and strict information (Chap. 2). Strict information is essential for representing definitional or taxonomic information (e.g., "all men are mortal", "penguins are birds"), and incorporating such information into the

---

[4]For a different probabilistic interpretation of the default rules, see Neufeld and Poole [93]. For a statistical interpretation, see the Bacchus [6].

4

knowledge base requires nontrivial changes in the notions of both consistency and entailment. This extension cannot be accomplished by simply treating to the strict conditional $\phi \Rightarrow \sigma$ as the material implication $\phi \supset \sigma$. For example, whereas the pair $\{b \supset f,\ b \supset \neg f\}$ is logically consistent, the desired semantics should render the set $\{b \Rightarrow f,\ b \Rightarrow \neg f\}$ inconsistent. In Chapter 2, I provide a probabilistic semantics for strict conditionals $\phi \Rightarrow \sigma$ as constraints on admissible probability functions forcing the conditional probability of σ given ϕ to be equal to 1. I then establish effective decision procedures for testing both consistency and entailment in knowledge bases containing mixtures of defeasible and strict information. Procedures for reasoning with inconsistent knowledge bases and ways of uncovering the set of rules responsible for the inconsistency are also examined.[5]

The second contribution is the formalization and characterization of more powerful notions of entailment (Chaps. 3 and 4). Default reasoning requires two facilities: One forcing retraction of conclusions in light of new refuting evidence (e.g., once we learn that the battery is dead, we no longer expect the engine to start); the other protecting conclusions from retraction in light of new but irrelevant evidence (e.g., the color of the car should not affect inferences regarding ignition keys, batteries, or engines). p-Entailment excels in the first task, but fails on the second because it is extremely cautious; it only sanctions conclusions that attain high probability in *all* probability distributions p-consistent with the knowledge base. In order to respect the communication convention that, unless stated explicitly, properties are presumed to be *irrelevant* to each other, we must consider only distributions that minimize dependencies, that is, they contain only the dependencies that are absolutely implied by the knowledge base (and none others).

Chapter 3 details an extension of p-entailment where dependencies are minimized via the principle of maximum entropy.[6] Chapter 3 also provides symbolic procedures for answering queries based on this principle, without the explicit computation of the maximum entropy distribution. A second extension of p-entailment, called system-Z^+, is presented in Chapter 4. System-Z^+ restricts the set of probability distributions to those that assign to each model the highest possible likelihood consistent with the default rules. The behavior of these two formalisms is compared and new insights on how semantical features influence the plausibility of the resulting theories are discussed. The approach based on maximum entropy yields more intuitive conclusions in some domains, but system-Z^+

---

[5]The results in this chapter were originally reported in Goldszmidt and Pearl [49].

[6]The use of maximum entropy in default reasoning as an extension of p-entailment was proposed by Pearl in [97].

provides considerable computational advantages. An earlier version of Chapter 3 can be found in Goldszmidt et. al. [47], while preliminary versions of the results in Chapter 4 were first reported in Goldszmidt and Pearl [50, 53]. These two formalisms, maximum entropy and system-Z^+, are completely independent of each other, and Chapter 3 is not a prerequisite for the understanding of Chapter 4.

The third contribution of this thesis is the development of a semantical theory and a computational facility for reasoning with variable strength defaults (Sec. 4.3) and soft or imprecise evidence (Sec. 4.5). The capability to reason with variable strength defaults is necessary in domains such as diagnosis, where the analyst may feel strongly that failures are more likely to occur in one type of devices (e.g., multipliers) than in other (e.g., adders). The capability of processing soft evidence is important when the context φ (of a query) is not given with absolute certainty, that is, when there is some vague testimony supporting φ but that testimony is undisclosed (or cannot be articulated using the basic propositions in our language, e.g., testimony of the senses) so that only a summary of that testimony saying that "φ is supported to a degree n" can be ascertained.

The introduction of graded defaults and soft evidence requires new query answering machinery which, in the traditional probabilistic setting turned out to be intractable.[7] This thesis shows that the symbolic nature of system-Z^+ admits a more manageable class of procedures; they require a polynomial number of propositional satisfiability tests and are therefore tractable for Horn expressions.

Augmenting the proposed semantics with the capability to represent causal relations, actions, and reasoning about change is the final contribution of this thesis (Chap. 5). This is accomplished by invoking the principle of *Markov shielding*, which imposes a stratified set of (probabilistic) independences among events. Informally, the principle can be stated as follows:

> Knowing the set of causes for a given effect renders the effect *independent* of all prior events.

I show how the incorporation of this principle gives rise to a norm of consistency, applicable to knowledge bases representing causal relations,[8] and how it solves some of the common problems associated with tasks of prediction and explanation reported in the nonmonotonic literature.

---

[7]Most logic-based schemes for default reasoning were also shown to be highly intractable [55] residing in the Σ_2^p level of the complexity hierarchy, compared with Δ_2^p in our system.

[8]To the best of my knowledge, this is the first consistency criterion devised to ensure the coherence of causal theories.

In Section 5.4, I demonstrate how the framework proposed in this dissertation can embody and unify the theories of belief revision (see Alchourrón, Gärdenfors and Makinson [3]) and belief updating (see Katsuno and Mendelzon [65]), two theories of belief change that have been developed independently of research in default reasoning and causal reasoning. Basically, theories of belief change seek general principles for constraining the process by which a rational agent ought to incorporate a new piece of information ϕ into an existing set of beliefs ψ, regardless of how the two are represented and manipulated. Belief revision deals with information obtained through new observations in a static world, while belief update deals with tracing changes in an evolving world (subjected perhaps to the external influence of actions).[9] I show that both revision and update can be modeled within the same framework using a qualitative version of probabilistic conditioning.

Finally, in Chapter 6 I discuss some open problems and suggest further challenges.

1.2 Extensional and Conditional Approaches

Approaches for formalizing defeasible reasoning can be loosely categorized as either *extensional* or *conditional*, depending on the interpretation assigned to the rule $\varphi \rightarrow \psi$.[10] Extensional approaches are based on "extending" classical logic by using defaults as rules for augmenting the sets of beliefs in the absence of conflicting evidence (see [87, 108, 89, 90, 45, 109, 38]). These approaches regard the default "if φ then ψ" as a qualified license believe ψ given the truth of φ. Conditional approaches, on the other hand, interpret the same rule as a *hard* but context dependent constraint to prefer ψ over $\neg\psi$ when φ is all that is known (see [36, 38, 69, 74, 23, 101, 14, 49, 50]). Conditional approaches are generally related to conditional logics studied in philosophy.[11]

As mentioned above, extensional approaches produced systems that exhibit many aspects of nonmonotonicity, thus allowing the retraction of conclusions in light of new information. However, they yield ambiguous results when confronted with conflicting defaults, with no way of distinguishing the intended from the unintended conclusions (see [112, 57]). In order to impose preferences and prevent generation of undesired inferences, special mechanisms must be devised to permit

---

[9]Preliminary versions of this chapter can be found in Goldszmidt and Pearl [54, 52].

[10]Not all formalisms can be categorized as either extensional or conditional. The approach based on multivalued logics proposed by Ginsberg [46] is one such example.

[11]For a survey of conditional logics see [94].

the specification of such preferences in the extensional approaches. These include, for example, special nonnormal defaults in default logic [30] (see Sec. 1.2.1) and priority-driven minimizations in circumscription [88] (see See. 1.2.2). Such mechanisms conflict with the original intent of default inference systems, since they require a possibly exhaustive enumeration of the exceptions to each default rule and/or an omniscient user capable of predicting and prioritizing all conceivable interactions among default rules.

In contrast, proposals based on conditional approaches have proven successful in enforcing the desired preferences in cases of conflicting defaults (see Examples 2.1 and 2.2). These preferences stem automatically from the semantical interpretation of the default rules, based on either an infinitesimal abstraction of probability theory or in *rankings* among possible worlds.[12] Unfortunately, the initial versions of these conditional formalisms failed to sanction some desirable patterns of inference that are readily sanctioned in common discourse [36, 97]. Their greatest limitation stems from the failure to properly handle irrelevant information. Recent extensions such as Delgrande's [23], Lehmann and Magidor's rational closure [74], and Pearl's system-Z [100], were successful in capturing some aspects of irrelevance, but are still unable to handle some cases, for example, property inheritance across exceptional subclasses (see Chapter 3). Solving these problems is one of the main contributions of this dissertation (Chaps. 3 and 4). Comparisons to Lehmann and Magidor's work can be found in Sections 3.2, 3.4, and 3.6.[13] System-Z is a special case of system-Z^+ developed in Chapter 4. Other conditional approaches described in the literature are Geffner's *conditional entailment* [36, 38] and Boutilier's modal logic CO* [14]. Conditional entailment is one of the most powerful formalisms for closing the *gap* between conditional and extensional approaches. It is reviewed in Section 4.7. Boutilier proves the equivalence between CO* and the notions of p-consistency and p-entailment [14]. He also axiomatizes system-Z in terms of Levesque's notion of *only knowing* formulation [75], and proves an interesting relation between the rule priorities of system-Z and the epistemic entrenchment of AGM [3, 33] (see Sec. 4.6).

Reiter's default logic [108], McCarthy's circumscription [87, 88], and Moore's autoepistemic logic [90] are reviewed next.[14] These three extensional approaches constituted the state of the art in the field when I began this research project.

---

[12]As it turns out these interpretations are practically equivalent (see [69] and also Chapter 3).

[13]Delgrande's [23] work is compared to Lehmann and Magidor's in [69].

[14]The descriptions in Secs. 1.2.1, 1.2.2, and 1.2.3 are not to be taken as detailed accounts of these formalisms. The reader is encouraged to examine the surveys in [45, 109] and to consult the relevant papers listed.

This review should highlight the parameters researchers use to judge progress in the field and clarify the significance of the contributions of this dissertation.

1.2.1 Reiter's Default Logic

The extension proposed by Reiter [108] is based on augmenting classical first order logic with inference rules of the form

$$\frac{\alpha(x) : \beta(x)}{\gamma(x)}, \tag{1.1}$$

where $\alpha(x), \beta(x)$, and $\gamma(x)$ are well-formed formulas (wffs) with free variables among those in x. The formula $\alpha(x)$ is called the precondition, $\beta(x)$ is called the test condition, and $\gamma(x)$ is the consequent of the default. Given a tuple of ground terms a, the rule in Eq. 1.1 allows us to conclude $\gamma(a)$ given that $\alpha(a)$ is believed and provided that $\beta(a)$ is *consistent* with the current set of beliefs. A default theory $T = \langle W, D \rangle$ is composed of a set W of wffs and a set D of default rules of the form specified by Eq. 1.1. Thus, given the theory

$$T = \langle \{bird(Tweety)\}, \{\frac{bird(x) : flies(x)}{flies(x)}\}\rangle, \tag{1.2}$$

we can derive $flies(Tweety)$. However, if $\neg flies(Tweety)$ can be established, for example, by augmenting W with

$$dead(Tweety) \supset \neg flies(Tweety), \text{ and } dead(Tweety),$$

the rule is blocked, and $flies(Tweety)$ is no longer a conclusion. Thus, nonmonotonicity is achieved by means of the *consistency check* required by the rules. Note that different rules also interact throughout this consistency check. For example, the theory

$$T = \langle \{penguin(Tweety), penguin(x) \supset bird(x)\},$$
$$\{\frac{penguin(x) : \neg flies(x)}{\neg flies(x)}, \frac{bird(x) : flies(x)}{flies(x)}\}\rangle \tag{1.3}$$

yields two possible *extensions*: one in which the first rule is blocked and *Tweety* flies, and the other in which the second rule is blocked and *Tweety* does not fly. Extensions are formally defined as follows: Let us say that $\Gamma(S)$ expands a set of wffs S according to T if $\Gamma(S)$ denotes the minimal deductively closed set of wffs which includes W and every consequent γ of default rules of the form $\alpha : \neg \beta / \gamma$ in D for which $\alpha \in \Gamma(S)$ and $\beta \neg \in S$. An extension of T is a collection E of wffs such that $E = \Gamma(E)$.

A default theory can give rise to one, none, or many extensions, and each extension is intended to reflect a possible completion of the classical theory W according to the rules in D. The natural encoding of a body of knowledge in the form of a default theory often gives rise to unreasonable extensions, which must be pruned (usually by the user) by properly selecting the test conditions of the defaults [112, 30, 29]. Thus, for example, in Eq. 1.3 the second default rule can be changed to read

$$\frac{bird(x) : flies(x) \wedge \neg penguin(x)}{flies(x)}, \tag{1.4}$$

and in general, the test condition should enumerate all anticipated exceptions. Default rules (such as the one in Eq. 1.4) in which the test condition is not equal to the consequent are commonly known as *nonnormal defaults* [30, 29].

On the positive side, Reiter's default logic extends classical first-order logic with nonmonotonic capabilities by means of a formal yet simple device, that is, by treating default rules as special rules of inference. Of all the extensional approaches, default logic appears to be the most stable: most work on default logic focuses on applying rather than modifying Reiter's original ideas [45]. Recent work extending default logic and solving some of its shortcomings can be found in [17, 24, 43]. Work on the computational complexity of default logic is reported in [66, 10], and the relation between default logic and formal semantics for logic programming is studied in [42, 11]. Default logic is compared to ε-semantics (a conditional approach underlying the development of Chap. 2) in [97].

1.2.2 McCarthy's Circumscription

Circumscription minimizes the extensions of various predicates in a given theory, thereby providing a closed world view of their interpretations. This formalism is best understood from a model-theoretic perspective. Let $A(P)$ denote a first order sentence containing the predicate P. In classical logic, a wff ψ is said to be entailed by $A(P)$ if ψ is true in every model for $A(P)$. Circumscription weakens this condition: ψ is entailed by $Circ[A(P); P]$ (to read: "the circumscription of P in $A(P)$") if ψ is true in every model of $A(P)$ which is *minimal* in P [88, 78]. A model M is minimal in P when there is no other model that assigns a strictly smaller extension to P and that preserves from M the same domain and the same interpretation of symbols other than P. Thus, given a set of axioms, circumscription selects a minimal interpretation for some predicate(s) subject to the constraints imposed by the axioms. As the set of axioms changes, so does the minimal interpretation that circumscription selects, and consequently

the set of inferred conclusions can shrink as new information arrives and the desired property of nonmonotonicity is attained. For instance, given a knowledge base containing the fact $penguin(Tweety)$, the circumscription of $penguin$ yields the formula $\forall x.penguin(x) \supset x = Tweety$. If $Opus$ is an object different from $Tweety$, circumscription will allow us to jump to the conclusion that $\neg penguin(Opus)$. If $penguin(Opus)$ is learned, the circumscription of $penguin$ will now yields $\forall x.penguin(x) \supset (x = Tweety \vee x = Opus)$.

Syntactically, the circumscription $Circ[A(P); P]$ of P in $A(P)$ can be expressed as the second-order schema [87]

$$A(P) \wedge A(\Phi) \wedge \forall x.[\Phi(x) \supset P(x)] \supset \forall x.[P(x) \supset \Phi(x)], \tag{1.5}$$

where $A(\Phi)$ denotes the logical sentence that results from replacing all the occurrences of P by a predicate Φ with the same arity as P. Eq. 1.5 can be understood as stating that among the predicates Φ that satisfy the constraints in $A(\Phi)$, P is the strongest; in other words, the objects that satisfy a predicate P are exactly the objects that can be shown to satisfy P.

Circumscription adds nonmonotonic features to first order logic but does not specify how defeasible knowledge should be encoded. McCarthy [88] introduced a convention by which defaults such as "birds fly" are written as

$$\forall x.bird(x) \wedge \neg ab_i(x) \supset flies(x) \tag{1.6}$$

and read as "every non-abnormal bird flies". Thus, given a set of these defaults, the expected behavior follows from minimizing the abnormalities, that is, from circumscribing the ab predicates. Note, however, that given Eq. 1.6 and $bird(Tweety)$, the minimization of ab_i by Eq. 1.5 will not suffice to sanction $flies(Tweety)$. This happens because the model in which *no bird is abnormal* and therefore $Tweety$ flies is competing with a model M' in which $ab_i(Tweety)$ and $\neg flies(Tweety)$ and M' is also minimal with respect to ab_i if we leave all the other objects constant. To remedy this undesirable situation, McCarthy [88] proposed a more powerful formula circumscription in which certain other predicates are allowed to vary, thus allowing the minimization of some predicates at the expense of others. The circumscription $Circ[A(P, Z) : P, Z]$ of the predicate P in $A(P, Z)$, where Z stands for a tuple of predicates allowed to vary in the minimization of P, is defined as

$$A(P, Z) \wedge A(\Phi, \Psi) \wedge \forall x.[\Phi(x) \supset P(x)] \supset \forall x.[P(x) \supset \Phi(x)] \tag{1.7}$$

Note that Eq. 1.7 is stronger than the schema in Eq. 1.5, since, in addition to substitutions for P, Eq. 1.7 permits substitutions for Z. The model-theoretic

11

interpretation of $Circ[A(P, Z); P, Z]$ sanctions as theorems the sentences that hold in all models for $A(P, Z)$ that are minimal in P with respect to Z [78]. A model M of $A(P, Z)$ is minimal in P with respect to Z, if there are no other models M' of $A(P, Z)$ that assign a smaller extension to P and that preserve from M the same domain and the same interpretation of symbols other than P and Z. Note that the expected conclusion, $flies(Tweety)$, follows in the example above by minimizing the ab_i predicate while allowing $flies$ to vary since the only minimal models are those in which $\neg ab_i(Tweety)$ holds.

The generalization of circumscription to the case of many predicates (known as *parallel circumscription*) is straightforward. A more interesting extension is that of *prioritized circumscription*, in which the user is allowed to specify a priority ordering among the predicates to be circumscribed, where predicates with higher priority are circumscribed (minimized) at the expense of predicates with lower priority [88, 81]. Thus, for example, if we add to Eq. 1.6

$$\forall x. penguin(x) \wedge \neg ab_j(x) \supset \neg flies(x) \tag{1.8}$$

$$\forall x. penguin(x) \supset bird(x) \tag{1.9}$$

$$penguin(Tweety) \tag{1.10}$$

then $\neg fly(Tweety)$ will follow only if we circumscribe ab_j with a higher priority than ab_i. Note that the *circumscriptive policy* – namely, the predicates to be minimized, the priority ordering, and the predicates to be allowed to vary – must be specified by the user.

Circumscription has been extensively studied due to its power and mathematical tractability. Circumscription shares some of the shortcomings of default logic: The user remains responsible for establishing preferences among default rules and for sorting out their possible interactions. Circumscription uses priorities among predicates on the minimization process to express such preferences. Lifschitz [80] reports on ways to incorporate the specification of such priorities into the object language. Efforts directed toward providing guidelines for specific domains can be found in [82, 8, 70]

1.2.3   Moore's Autoepistemic Logic

Moore [90] originally proposed autoepistemic logic as a reconstruction of McDermott and Doyle's nonmonotonic logic [89]. Autoepistemic logic augments propositional theories with a belief operator L, where sentences of the form $L\varphi$ are read as "φ is believed". The *stable expansion* of an autoepistemic theory T,

$S(T)$, is defined as follows

$$S(T) = Th(T \cup \{Lp : p \in S(T)\} \cup \{\neg Lp : p \notin S(T)\}) \qquad (1.11)$$

where $Th(X)$ stands for the set of tautological consequences of X. Stable expansions are intended to reflect possible states of belief of an ideal rational agent, *closed* under both negative and positive introspection [90].

Defaults can be encoded in autoepistemic logic using an *ab* predicate similar to circumscription; thus, "typically birds fly" will be written $bird \wedge \neg Lab_i \supset flies$. Given *bird*, the only autoepistemic expansion will contain $\neg Lab_i$ and consequently the proposition *flies*. An autoepistemic theory may have one, none, or many stable expansions. For instance, $T = \{\neg Lp \supset p\}$ has no stable expansion, while $T = \{\neg Lp \supset q, \neg Lq \supset p\}$ has two.

Since its introduction, autoepistemic logic has been studied by [86, 67, 41, 91]. It has been successfully applied to characterize the semantics of general logic programs [40, 42] and of truth maintenance systems [107]. Both characterizations require only the replacement of logical negation by autoepistemic negation, that is, literals of the form $\neg p$ are replaced by $\neg Lp$. Levesque [75] provides an appealing semantics for autoepistemic logic in terms of *only knowing* (see also [14]).

As in the case of default logic and circumscription, autoepistemic logic is unable to automatically account for preferences among defaults and resolve their interactions in a satisfactory manner. As we shall see, this problem is solved in this dissertation by interpretating default rules as preference constraints on the set of possible situations. The basis for this interpretation is a norm of consistency to be introduced next in Chapter 2.

CHAPTER 2

The Consistency of Conditional Knowledge Bases

2.1 Introduction

There is a sharp difference between exceptions and outright contradictions. The two statements "typically penguins do not fly" and "red penguins fly" can be accepted as a description of a world in which *redness* defines an abnormal or exceptional type of penguin. However, the statements s_1:"typically birds fly" and s_2:"typically birds do not fly" stand in outright contradiction to each other. Whatever interpretation we give to "typically", it is hard to imagine a *world* containing birds in which both s_1 and s_2 would make sense simultaneously. Curiously, such conflicting pairs of sentences can coexist perfectly in most *nonmonotonic* formalisms directed at capturing and characterizing our everyday reasoning by including such expressions about what is normally the case. For example, using the *ab* predicate advocated by McCarthy [88], a straightforward way to represent such statements in the context of *circumscription* would be

$$s_1' : \forall x.\mathsf{bird}(x) \wedge \neg ab(x) \supset \mathsf{fly}(x) \quad ; \quad s_2' : \forall x.\mathsf{bird}(x) \wedge \neg ab(x) \supset \neg\mathsf{fly}(x), \quad (2.1)$$

which is logically equivalent to $\forall x.\mathsf{bird}(x) \supset ab(x)$. Similarly, if s_1 and s_2 are expressed as the *default rules*[1]

$$s_1'' : \frac{\mathsf{bird}(x) : \mathsf{M} \; \mathsf{fly}(x)}{\mathsf{fly}(x)} \quad ; \quad s_2'' : \frac{\mathsf{bird}(x) : \mathsf{M} \; \neg \; \mathsf{fly}(x)}{\neg \; \mathsf{fly}(x)}, \quad (2.2)$$

Reiter's default logic [108] will produce two consistent sets of beliefs, one in which "birds fly" and one in which "birds do not fly".

Normally, a pair such as s_1 and s_2 would not be used to encode the information that "all birds are *exceptional* (or *abnormal*)" as in the case of circumscription or to express an *ambiguous* property[2] of birds as in the case of default logic.

---

[1] The default rule $\frac{\mathsf{bird}(x):\mathsf{M} \; \mathsf{fly}(x)}{\mathsf{fly}(x)}$ is informally interpreted as "if x is a bird and it is consistent to assume that x can fly, then infer that x can fly" (see [108]).

[2] A property f is ambiguous if neither f nor $\neg f$ can be verified from the knowledge base.

Rather, this kind of contradictory information is more likely to originate from an *unintentional* mistake. Remarkably, although humans readily recognize the distinction between exceptions, ambiguities, and contradictions, current work on defeasible knowledge bases presents no comprehensive analysis of such utterances, which could alert the user to the existence of contradictory, possibly unintended statements. As a first step in formulating a framework for representing and reasoning with if–then rules admitting exceptions, this chapter proposes a semantically sound norm for consistency, accompanied by effective procedures for testing inconsistencies and isolating their origins.

It is tempting to assume that pairs such as s_1 and s_2 constitute the only source of inconsistency and that once we eliminate such contradictory pairs, the remaining knowledge base would be consistent, that is, all conflicts could be rationalized as conveying exceptions or ambiguities. Touretzky [122] has shown that this is indeed the case in the domain of acyclic and purely defeasible *inheritance networks*. However, once the language becomes more expressive, allowing hard rules as well as arbitrary formulas in the antecedents and consequents of the rules, the criterion for consistency becomes more involved. Consider the knowledge base $\Delta =$ {"all birds fly", "typically penguins are birds", "typically penguins do not fly"}. This set of rules, although without contradictory pairs, also strikes us as inconsistent: If all birds fly, there cannot be a nonempty class of objects (penguins) that are "typically birds" and yet "typically do not fly". We cannot accept this knowledge base as merely depicting exceptions; it looks more like a programming "bug" or "glitch" than a genuine description of some state of affairs. If we now change the first sentence to read "typically birds fly" (instead of "all birds fly"), consistency is restored; we are willing to accept penguins as exceptional birds. This interpretation would remain satisfactory even if we made the second rule strict (to read "all penguins are birds"). Yet, if we add to Δ the sentence "typically birds are penguins", we again face intuitive *inconsistency*.

In this chapter we propose a probability-based formalism that captures these intuitions. We will interpret a *defeasible* rule "typically, if φ then ψ" (written $\varphi \rightarrow \psi$) as the conditional probability statement $P(\psi|\varphi) \geq 1 - \varepsilon$, where $\varepsilon > 0$ is an infinitesimal quantity. Intuitively, this amounts to according the consequence ψ a very high likelihood whenever the antecedent φ is all that we know. The *strict* rule "if ϕ then definitely σ" (written $\phi \Rightarrow \sigma$) will be interpreted as an extreme conditional probability statement $P(\sigma|\phi) = 1$. Our criterion for testing consistency translates to determining whether there exists a probability distribution P that satisfies all these conditional probabilities for every $\varepsilon > 0$. Furthermore, to match our intuition that conditional rules neither refer to empty classes nor are confirmed by merely "falsifying" their antecedents, we also require that P be

proper, that is, it does not render any antecedent as totally impossible. These two requirements constitute the essence of our proposal.

In the language of ranked models (see Secs. 3.2, and 4.2, and also [72]), our proposal assumes a particularly simple form. A defeasible rule $\varphi \rightarrow \psi$ imposes the constraints that ψ holds in all minimally ranked models of φ and that there will be at least one such model. A strict rule $\phi \Rightarrow \sigma$ imposes the constraint that no possible world satisfies $\phi \wedge \neg\sigma$ and that at least one possible world satisfies ϕ. Consistency amounts to requiring the existence of a ranking (a mapping of models to integers) that simultaneously satisfies all these constraints. The idea of attaching probabilistic semantics to conditional rules goes back to Adams [1, 2], who developed a logic of indicative conditionals based on infinitesimal probabilities.[3] More recently, infinitesimal probabilities were mentioned by McCarthy [88] as a possible interpretation of circumscription and were used by Pearl [95] to develop a graphical consistency test for inheritance networks, extending that of Touretzky [122]. The proposals in [97, 102, 35, 37] have extended Adams' logic to default schemata, and Lehmann and Magidor [74] have shown the equivalence between Adams' logic and a semantics based on ranked models.

Unfortunately, the notion of consistency treated in [2] and [95] was restricted to systems involving purely defeasible rules. This chapter extends Adams' consistency results to mixed systems containing both defeasible and strict information, and, as we shall see, the extension is by no means trivial, since a strict rule $b \Rightarrow f$ must be given a semantics totally different from its material counterpart $b \supset f$. For example, whereas the set of rules $\{b \supset f, b \supset \neg f\}$ is logically consistent, our semantics must now render the set $\{b \Rightarrow f, b \Rightarrow \neg f\}$ inconsistent. The need to distinguish between $b \Rightarrow f$ and $b \supset f$, where the former is used to express generic knowledge and the latter as an item of evidence is also advocated in [35, 37, 23, 104] (see Sec. 2.7). The implications of this distinction will become more apparent in Chapter 5, where causality is introduced in the interpretation of the conditional rules.

In addition to extending the consistency criterion to include mixed systems, we also present an effective syntactic procedure for testing this criterion and identifying the set of rules responsible for the inconsistency. Finally, we analyze a notion of entailment based on consistency considerations. Intuitively, a conclusion is entailed by a knowledge base if it is guaranteed an arbitrarily high probability whenever the premises are assigned sufficiently high probabilities. This weak notion of entailment was named p-entailment in [2], ε-entailment in [97],

---

[3] A formal treatment of infinitesimal probabilities using nonstandard analysis is given in [74] and also mentioned in [120].

and preferential entailment in [69], and it yields (semimonotonically) the most conservative "core" of plausible conclusions that one would wish to draw from a conditional knowledge base [98].

The definition for probabilistic entailment can be partially extended to knowledge bases containing strict information using a device suggested by Adams [1] whereby, by definition, conditional rules whose antecedents have probability zero are assigned probability one. Thus, a strict rule such as $\phi \Rightarrow \sigma$ one could conceivably be encoded as the defeasible rule $(\phi \wedge \neg\sigma) \rightarrow False$. Another proposal was made in the preferential-models analysis of [69]. There, Kraus, Lehmann, and Magidor write (p. 172):

> We reserve to ourselves the right to consider universes of reference that are strict subsets of the sets of all models of L. In this way, we shall be able to model *strict* constraints, such as *penguins are birds*, in a simple and natural way, by restricting $\mathcal{U}$ to the set of all worlds that satisfy the material implication *penguin* $\supset$ *bird*.

Both of these proposals suffer from two weaknesses. First, they do not capture the common understanding that the opposing pair "all birds fly" and "all birds don't fly" is inconsistent, but instead permit the conclusion that birds do not exist, together with other strange consequences such as "typically birds have property P" where P stands for any imaginable property. Our semantics reflects the view, also expressed in [23], that one of the previous rules must be invalid and that no admissible model would support both rules. Second, these proposals do not permit us to entail new strict rules in a more meaningful way, according to our commonsense interpretation of conditional sentences, than logical entailment. For example, $\neg a$ should not entail $a \Rightarrow b$, in the same way that "I am poor" should not entail "if I were rich, it should rain tomorrow". Thus, the special semantics we give to conditional rules, defeasible as well as strict, avoids such paradoxes of material implication [4] and, hence, brings mechanical and plausible reasoning closer together.

This chapter is organized as follows: Section 2.2 introduces notation and some preliminary definitions. Consistency and entailment are explored in Section 2.3. An effective procedure for testing consistency and entailment is presented in Section 2.4, while Section 2.5 contains illustrative examples. Section 2.6 deals with entailment in inconsistent knowledge bases, and the main results are summarized in Section 2.7. All proofs appear in Appendix A.

2.2 Notation and Preliminary Definitions

The basic language is a finite set $\mathcal{L}$ of atomic propositions augmented with two propositional constants T and F, which are (informally) regarded as expressing a logical truth and a logical falsehood, respectively. Let $\mathcal{L}_P$ be a closed set of propositional well-formed formulas (wffs) generated as usual from the atomic propositions in $\mathcal{L}$ and the connectives $\vee$ and $\neg$. We define a *world* ω as a truth assignment for the atomic propositions in $\mathcal{L}$. The set of possible worlds is denoted by Ω, and if there are n atomic propositions in $\mathcal{L}$, the size of Ω will be 2^n. The satisfaction of a wff $\varphi \in \mathcal{L}_P$ by a world ω is defined as usual and denoted by $\omega \models \varphi$. If ω satisfies φ, we say that ω is a *model* for φ.

A *defeasible rule* is the formula $\varphi \rightarrow \psi$, where φ and ψ are wffs in $\mathcal{L}_P$ and $\rightarrow$ is a new binary connective. Informally, each $\varphi \rightarrow \psi$ represents an if–then rule that admits exceptions and each may be read as "if φ then *typically* ψ" or "if φ then *normally* ψ". Similarly, given ϕ, σ in $\mathcal{L}_P$, the new binary connective $\Rightarrow$ will be used to form a *strict rule* $\phi \Rightarrow \sigma$. A strict rule $\phi \Rightarrow \sigma$ is interpreted as "if ϕ then *definitely* σ". A formal interpretation of both strict and defeasible rules is given in the definition of consistency (Def. 2.2). Both $\rightarrow$ and $\Rightarrow$ can occur only as the main connective in a rule. We will use *conditional rules* or simply *rules* when referring to a formula that can be either a defeasible or a strict rule. The *antecedent* of a rule is the wff to the left of the main connective (single or double arrow) and its *consequent* is the wff to the right. If r denotes a conditional rule with antecedent ϕ and consequent ψ, then the *negation* of r, denoted by $\sim r$, is defined as a conditional with antecedent ϕ and consequent $\neg\psi$. The *material counterpart* of a conditional rule with antecedent φ and consequent ψ is defined as $\varphi \supset \psi$ (where $\supset$ denotes material implication), and the material counterpart of a set Δ of conditional rules (denoted by $\hat{\Delta}$) is defined as the conjunction of the material counterparts of the rules in Δ.

A default $\varphi \rightarrow \psi$ is *verified* by a world ω iff $\omega \models \varphi \wedge \psi$. $\varphi \rightarrow \psi$ is *falsified* by ω iff $\omega \models \varphi \wedge \neg\psi$. Finally, $\varphi \rightarrow \psi$ is *satisfied* by ω iff $\omega \models \varphi \supset \psi$. Strict rules are verified, falsified, and satisfied in the same way.

Definition 2.1 (Probability assignment) Let P be a probability function on the space of possible worlds Ω, such that $P(\omega) \geq 0$ and $\sum_{\omega \in \Omega} P(\omega) = 1$. We define a *probability assignment* P on a formula $\varphi \in \mathcal{L}$ as

$$P(\varphi) = \sum_{\omega \models \varphi} P(\omega). \tag{2.3}$$

Let $\Delta = D \cup S$ be a set of conditional rules such that $D = \{\varphi_i \rightarrow \psi_i\}$ $(1 \leq i \leq |D|)$ and $S = \{\phi_j \Rightarrow \sigma_j\}$ $(1 \leq j \leq |S|)$. A probability assignment on a defeasible rule

18

$\varphi \rightarrow \psi \in D$ is defined as

$$P(\varphi \rightarrow \psi) = \begin{cases} \frac{P(\varphi \wedge \psi)}{P(\varphi)} = P(\psi|\varphi) & \text{if } P(\varphi) > 0 \\ 1 & \text{otherwise} \end{cases} \tag{2.4}$$

We assign probabilities to the rules in $\mathcal{S}$ in exactly the same fashion. P will be considered *proper* for a conditional rule r with antecedent φ if $P(\varphi) > 0$, and it will be proper for Δ if it is proper for every conditional in Δ.
□

The probability assignment above attaches a conditional probability interpretation to the rules in a given Δ. Eq. 2.4 states that the probability of a conditional rule r with antecedent φ and consequent ψ is equal to the probability of r being verified (i.e., $\omega \models \varphi \wedge \psi$) divided by the probability of its being either verified or falsified (i.e., $\omega \models \varphi$).

Up to this point the only difference between defeasible and strict rules is syntactic. They are assigned probabilities in the same fashion and are verified and falsified under the same truth assignments. Their differences will become clear in the next section, where we formally introduce the notion of *consistency*.

2.3   Probabilistic Consistency and Entailment

Throughout the rest of the chapter, Δ denotes a knowledge base of conditional rules. $\Delta = D \cup S$, where $D = \{\varphi_i \rightarrow \psi_i\}$ ($1 \leq i \leq |D|$) and $S = \{\phi_j \Rightarrow \sigma_j\}$ ($1 \leq j \leq |S|$).

Definition 2.2 (Probabilistic consistency) We say that $\Delta = D \cup S$ is *probabilistically consistent* (p-consistent) if for every $\varepsilon > 0$, there is a probability assignment P that is proper for Δ such that $P(\psi|\varphi) \geq 1 - \varepsilon$ for all defeasible rules $\varphi \rightarrow \psi$ in D and $P(\sigma|\phi) = 1$ for all strict rules $\phi \Rightarrow \sigma$ in S.
□

Intuitively, consistency means that it is possible for all defeasible rules to come as close to certainty as desired, while all strict rules hold with absolute certainty. Another way of formulating consistency is as follows: Consider a constant $\varepsilon > 0$ and let $\mathcal{P}_{\Delta,\varepsilon}$ stand for the set of proper probability assignments for Δ such that if $P \in \mathcal{P}_{\Delta,\varepsilon}$ then $P(\psi|\varphi) \geq 1 - \varepsilon$ for every $\varphi \rightarrow \psi \in D$ and $P(\sigma|\phi) = 1$ for every $\phi \Rightarrow \sigma \in S$. Consistency insists on $\mathcal{P}_{\Delta,\varepsilon}$ being nonempty for every $\varepsilon > 0$.

Before developing a syntactical test for consistency (Thm. 2.4), we need to define the concept of *toleration*.

Definition 2.3 (Toleration) Let r be a rule (either defeasible or strict) with antecedent α and consequent β. We say that r is *tolerated* by a set Δ if there exists a world ω such that

$$\omega \models \alpha \wedge \beta \bigwedge_{i=1}^{i=|D|} \varphi_i \supset \psi_i \bigwedge_{j=1}^{j=|S|} \phi_j \supset \sigma_j. \tag{2.5}$$

$\square$

Thus, r is tolerated by a set of conditional rules Δ if there is a world ω that verifies x and satisfies every rule in Δ (i.e., no rule in Δ is falsified by ω).

Theorem 2.4 *Let $\Delta = D \cup S$ be a nonempty set of defeasible and strict rules. Δ is p-consistent iff every nonempty subset $\Delta' = D' \cup S'$ of Δ complies with one of the following:*

1. *If D' is not empty, then there must be at least one defeasible rule in D' tolerated by Δ'.*

2. *If D' is empty (i.e., $\Delta' = S'$), each strict rule in S' must be tolerated by S'.*

The following corollary ensures that, in order to determine p-consistency, it is not necessary to check literally every nonempty subset of Δ.

Corollary 2.5 *$\Delta = D \cup S$ is p-consistent iff we can build an ordered partition of $D = [D_1, D_2, \ldots, D_n]$ where*

1. *For all $1 \leq i \leq n$, each rule in D_i is tolerated by $S \cup_{j=i+1}^{j=n} D_j$.*

2. *Every rule in S is tolerated by S.*

Corollary 2.5 reflects the following considerations (see proof in Appendix A): If Δ is p-consistent, Theorem 2.4 ensures the construction of the ordered partition. On the other hand, if this partition can be built, the proof of Theorem 2.4 shows that a probability assignment can be constructed to comply with the requirements of Def. 2.2. Corollary 2.5 yields a simple and effective decision procedure for determining p-consistency and identifying the inconsistent subset in Δ (see Sec. 2.4).

Before turning to the task of entailing new rules, we need to make explicit a particular form of inconsistency.

Definition 2.6 (Substantive inconsistency) Let Δ be a p-consistent set of conditional rules, and let r' be a conditional rule with antecedent ϕ. We will say that r' is *substantively inconsistent* with respect to Δ if $\Delta \cup \{\phi \to True\}$ is p-consistent but $\Delta \cup \{r'\}$ is p-inconsistent.
$\square$

Nonsubstantive inconsistency occurs whenever the antecedent of a conditional rule is logically incompatible with the strict rules of a consistent set Δ. It will become apparent from the theorems to follow that a rule r is nonsubstantively inconsistent with respect to a consistent Δ iff both $\Delta \cup \{r\}$ and $\Delta \cup \{\sim r\}$ are inconsistent.

The concept of *entailment* introduced below is based on the same probabilistic interpretation as the one used in the definition of p-consistency. Intuitively, we want p-entailed conclusions to receive arbitrarily high probability in every proper probability distribution in which the defeasible premises have sufficiently high probability and in which the strict premises have probability equal to one.

Definition 2.7 (p-Entailment) Given a p-consistent set Δ of conditional rules, Δ *p-entails* $\varphi' \to \psi'$ (written $\Delta \models_p \varphi' \to \psi'$) if for all $\varepsilon > 0$ there exists $\delta > 0$ such that

1. There exists at least one $P \in \mathcal{P}_{\Delta,\delta}$ [4] such that P is proper for $\varphi' \to \psi'$.

2. Every $P' \in \mathcal{P}_{\Delta,\delta}$ satisfies $P'(\psi'|\varphi') \geq 1 - \varepsilon$.

$\square$

Theorem 2.8 relates the notions of entailment and consistency.

Theorem 2.8 *If Δ is p-consistent, Δ p-entails $\varphi' \to \psi'$ iff $\phi' \to \neg\psi'$ is substantively inconsistent with respect to Δ.*

Def. 2.9 and Theorem 2.10 characterize the conditions under which conditional conclusions are guaranteed not only very high likelihood but also absolute certainty. We call this form of entailment *strict p-entailment*.

---

[4] Recall that given a consistent $\Delta = D \cup S$, $\mathcal{P}_{\Delta,\delta}$ stands for the set of probability assignments proper for Δ, such that if $P \in \mathcal{P}_{\Delta,\varepsilon}$ then $P(\psi|\varphi) \geq 1 - \delta$ for every $\varphi \to \psi \in D$ and $P(\sigma|\phi) = 1$ for every $\phi \Rightarrow \sigma \in S$ (see Def. 2.2).

Definition 2.9 (Strict p-entailment) Given a p-consistent set Δ of conditional rules, Δ *strictly p-entails* $\phi' \Rightarrow \sigma'$ (written $\Delta \models_s \phi' \Rightarrow \sigma'$) if for all $\varepsilon > 0$

1. There exists at least one $P \in \mathcal{P}_{\Delta,\varepsilon}$ such that P is proper for $\phi' \Rightarrow \sigma'$.

2. Every $P' \in \mathcal{P}_{\Delta,\varepsilon}$ satisfies $P'(\sigma'|\phi') = 1$.

$\square$

Theorem 2.10 *If $\Delta = D \cup S$ is p-consistent, Δ strictly p-entails $\phi' \Rightarrow \sigma'$ iff $S \cup \{\varphi' \rightarrow True\}$ is p-consistent and there exists a subset S' of S such that $\varphi' \Rightarrow \neg\sigma'$ is not tolerated by S'.*

Examples of strict p-entailment are contraposition, $\{\phi \Rightarrow \psi\} \models_s \neg\psi \Rightarrow \neg\phi$,[5] and chaining $\{\phi \Rightarrow \sigma, \sigma \Rightarrow \psi\} \models_s \phi \Rightarrow \psi$. Note that strict p-entailment subsumes p-entailment, that is, if a conditional rule is strictly p-entailed, then it is also p-entailed. Also, to test whether a conditional rule is strictly p-entailed, we need to check its status only with respect to the strict set in Δ. This confirms the intuition that we cannot deduce "hard" rules from "soft" ones.

Note that the requirements of substantive consistency in Theorem 2.10 and properness for the probability distributions in Definition 2.2 distinguish strict rules from their material counterparts and establish a difference between strict p-entailment and logical entailment. For example, consider the knowledge base $\Delta = S = \{c \Rightarrow \neg a\}$, which is clearly p-consistent. While $\hat{\Delta} = \{c \supset \neg a\}$ logically entails $c \wedge a \supset b$, Δ does not strictly p-entail $c \wedge a \Rightarrow b$, since the antecedent $c \wedge a$ is always falsified.

Theorems 2.11 and 2.12 present additional results relating consistency and entailment. They follow immediately from previous theorems and definitions. Versions of these theorems, for the case of knowledge bases containing only defeasible rules first appeared in [2].

Theorem 2.11 *If Δ does not p-entail $\varphi' \rightarrow \psi'$, and $\varphi' \rightarrow \psi'$ is substantively inconsistent with respect to Δ, then for all $\varepsilon > 0$ there exists a probability assignment $P' \in \mathcal{P}_{\Delta,\varepsilon}$ which is proper for Δ and $\varphi' \rightarrow \psi'$ such that $P'(\psi'|\varphi') \leq \varepsilon$.*

Theorem 2.12 *If $\Delta = D \cup S$ is p-consistent, then it cannot be the case that*

1. *Both $\varphi \rightarrow \psi$ and $\varphi \rightarrow \neg\psi$ are substantively inconsistent with respect to Δ.*

2. *Both $\phi \Rightarrow \sigma$ and $\phi \Rightarrow \neg\sigma$ are substantively inconsistent with respect to S.*

---

[5]Whenever $\neg\psi$ is satisfiable.

Procedure Test_Consistency
Input: A set of rules $\Delta = D \cup S$.
Output: Yes/No depending on whether R is consistent.

1. Let $D' := D$.

2. While D' is not empty, do:

    2.1 Find a rule $d : \varphi \rightarrow \psi \in D'$ such that d is tolerated by $S \cup D'$; let $D' := D' - d$.

    2.2 If d is not found, abort: Return(No), Δ is inconsistent.

3. Let $S' := S$.

4. While S' is not empty, do:

    4.1 Pick any rule $s : \phi \Rightarrow \sigma \in S'$; if s is tolerated by S, then let $S' := S' - s$.

    4.2 Else abort: Return(No), Δ is inconsistent.

5. Return(Yes), Δ is consistent.

End Procedure

Figure 2.1: An effective procedure for testing consistency in $O(|D|^2 + |S|)$ propositional satisfiability tests.

2.4   An Effective Procedure for Testing Consistency

In accordance with Theorem 2.4 and following Corollary 2.5, the consistency of $\Delta = D \cup S$ can be tested in two phases. In the first phase, until D is empty we repeatedly remove from D a defeasible rule that is tolerated by the rest of the rules in $D \cup S$. In the second phase, we test whether every strict rule in S is tolerated by the rest of S (without removing any rule). If both phases can be successfully completed, Δ is consistent; if not, Δ is inconsistent. Procedure Test_Consistency is formally presented in Figure 2.1.

The same procedure can be used for entailment, since to determine whether a defeasible rule d' is entailed by Δ we need only test the consistency of $\Delta \cup \{\sim d'\}$ and $\Delta \cup \{d'\}$ (to make sure that the former is substantively inconsistent). Given that the procedure in Figure 2.1 is a sound and complete test for deciding p-consistency, the next theorem establishes an upper bound for the problem of deciding p-consistency (and p-entailment). Theorem 2.13 and the correctness of the procedure Test_Consistency are proven in Appendix A.

Theorem 2.13 *The worst case complexity of testing consistency (or entailment) is bounded by $[\mathcal{PS} \times (\frac{|D|^2}{2} + |S|)]$, where $|D|$ and $|S|$ are the number of defeasible and strict rules, respectively, and $\mathcal{PS}$ is the complexity of testing propositional satisfiability for the material counterpart of the rules in the database.*

Thus, the complexity of deciding p-consistency and p-entailment is no worse than that of propositional satisfiability. Although the general satisfiability problem is NP-complete, useful sublanguages (e.g., Horn clauses) are known to admit polynomial algorithms [25].

The order in which rules are removed in procedure Test_Consistency induces natural priorities among defaults that have been used to great advantage in several proposals for default reasoning, as is shown in Chapter 4 (see also [50, 100, 47, 36]). These priorities have an alternative epistemic interpretation in the theory of belief revision described by Gärdenfors [33]. The fact that a conditional $\varphi \to \psi$ is tolerated by all those rules that were not previously removed from Δ means that if φ holds, then ψ can be asserted without violating any rule in Δ that is more deeply entrenched than this conditional. In other words, adding the assertion $\varphi \wedge \psi$ would require a minimal revision of the set of beliefs supported by Δ. The formal relation between the default priorities used in both system-Z [100] and system-Z^+ [50] (see Sec. 4.6) and the postulates for epistemic entrenchment in believe revision [33] is studied by Boutilier [13]. The origin of this priority ordering can be traced back to Adams [2], where it is used to build "nested sequences" subsets of Δ that yield consistent, high probability models. Such "nested sequences" are used in the proof of Theorem 2.4 (see Appendix A). A similar construction was also used in [72, Theorem 5] to prove the co-NP-completeness of p-entailment in the case of knowledge bases containing only defeasible rules.

Once a set of rules is found to be p-inconsistent, it would be useful to identify the rules that are *directly responsible* for the contradiction. Unfortunately, the toleration relation is not strong enough to accomplish this task since it is incapable of distinguishing a rule "causing" the inconsistency from one that is a "victim" of the inconsistency. For example, consider the inconsistent set $D_i = \{\phi \to \psi, \phi \to \neg\psi, \phi \to \sigma\}$. Since no rule in D_i is tolerated, the consistency test will immediately halt and declare D_i inconsistent. Yet $\phi \to \sigma$ can hardly be held responsible for the inconsistency; $\phi \to \sigma$ is not tolerated because the material counterpart of the pair $\{\phi \to \psi, \phi \to \neg\psi\}$ renders ϕ impossible.[6] It would be inappropriate to treat a rule as the source of inconsistency merely because it is not tolerated in the context of an unconfirmable subset. Rather, we would like

---

[6]Note that $\{\phi \supset \psi, \phi \supset \neg\psi\} \models \neg\phi$.

to proclaim a rule *inconsistent* if its removal would improve the consistency of the database. In other words, a conditional rule r is inconsistent with respect to a set Δ iff there is an inconsistent subset of Δ that becomes consistent after r is removed. Formally,

Definition 2.14 (Inconsistent rule) A rule r is *inconsistent with respect to a set* Δ iff there exists a subset Δ' of Δ such that $\Delta' \cup \{r\}$ is p-inconsistent but Δ' in itself is p-consistent.
$\square$

Deciding whether a given rule is inconsistent is difficult because, unlike the test for set inconsistency, the search for the indicative subset Δ' cannot be systematized as in procedure Test_Consistency. All indications are that the search for such a subset will require exponential time. Simple-minded procedures based on removing one rule at a time and testing for consistency in the remaining set do not yield the desired results. In $\Delta' = \{a \rightarrow b, a \rightarrow \neg b, a \rightarrow c, a \Rightarrow \neg c\}$ every rule is inconsistent, yet it is necessary to remove at least two rules at a time in order to render the remaining set consistent. Likewise, in $\Delta'' = \{a \rightarrow b, a \rightarrow \neg b, a \rightarrow c, c \Rightarrow \neg b\}$ every rule is inconsistent, yet only the removal of $a \rightarrow b$ renders the remaining set consistent (or confirmable). Approximate methods for identifying inconsistent rules are discussed in Section 2.6 and in the proof of Theorem 2.24 (see Appendix A).

2.5 Examples

The following examples depict some of the rule interactions commonly found in everyday discourse which motivated the development of nonmonotonic logics and formalisms for default reasoning. They represent benchmarks in nonmonotonic reasoning and will be used throughout the thesis. As a common denominator, Examples 2.1, 2.2, and 2.3 contain a pair of conflicting rules. Example 2.1 refers to the case of one if–then rule denoting what is generally the case, "if φ then ψ", and another if-then rule representing an exception (φ and γ) to the first one, "if φ and γ then $\neg\psi$". Example 2.2 is similar, except that the antecedents of the conflicting rules "if φ then ψ" and "if γ then $\neg\psi$" are related through a third rule, "if γ then φ", which points out that γ is a *more specific context* than α. Finally, the antecedents in the rules of Example 2.3 are unrelated. Thus, the conflict cannot be resolved and the conclusion remains ambiguous (i.e., neither ψ nor $\neg\psi$ is sanctioned). In all the examples, the rules are modified slightly to highlight the differences between exceptions, contradictions, and ambiguities.

Example 2.1 (Dead battery) Consider the following rules:

1. $t \Rightarrow c$ ("if I turn the ignition key, definitely the car will start").

2. $t \wedge b \rightarrow \neg c$ ("if I turn the key and the battery is dead, then normally the car will not start").

This knowledge base is p-inconsistent: Any world $\omega \models t \wedge b \wedge \neg c$ (verifying Rule 2, the only defeasible rule) will falsify Rule 1. Intuitively, if the car engine will always start when the ignition key is turned, we cannot accept any *faults* (e.g., a dead battery). By changing the first rule to be defeasible, we obtain a p-consistent knowledge base Δ_c:

1. $t \rightarrow c$ ("if I turn the ignition key, then normally the car will start").

2. $t \wedge b \rightarrow \neg c$ ("if I turn the key and the battery is dead, then normally the car will not start").

The first rule is tolerated by the second using any world $\omega \models t \wedge \neg b \wedge c$ (and once Rule 1 is removed, Rule 2 is trivially tolerated by the remaining empty set). Among the p-entailed conclusions, we have

1. $\Delta_c \models_p t \rightarrow c$ ("if I turn the ignition key, then normally the car will start").

2. $\Delta_c \models_p t \wedge b \rightarrow \neg c$ ("if I turn the ignition key and the battery is dead, then normally the car will not start").

3. $\Delta_c \models_p t \rightarrow \neg b$ ("normally, when I turn the ignition key the battery is not dead").

Example 2.2 (Penguins and birds) Consider the knowledge base presented in the introduction:

1. $b \Rightarrow f$ ("all birds fly").

2. $p \rightarrow b$ ("typically penguins are birds").

3. $p \rightarrow \neg f$ ("typically penguins don't fly").

Clearly, none of the defeasible rules in the example can be tolerated by the rest. Consider a world ω, such that $\omega \models p \wedge b$ (testing whether Rule 2 is tolerated). If $\omega \models f$ Rule 3 will be falsified, while if $\omega \models \neg f$ Rule 1 will be falsified. Thus, we conclude that there is no world such that Rule 2 is tolerated. A similar situation

arises when we check whether Rule 3 can be tolerated. Thus, this knowledge base is p-inconsistent. Making Rule 1 defeasible yields the so-called "penguin triangle", $D_p = \{b \to f, p \to b, p \to \neg f\}$, which is p-consistent: $b \to f$ is tolerated by Rules 2 and 3 through the world ω', where $\omega' \models b \wedge f$ and $\omega' \models \neg p$, and, once Rule 1 is removed, the remaining rules tolerate each other. D_p becomes p-inconsistent by adding the rule $b \to p$ ("typically birds are penguins"), in conformity with the graphical criterion of [95]. Note that, by Theorem 2.8, the rule $b \to \neg p$ ("typically birds are not penguins") is then p-entailed by D_p. To demonstrate an inconsistency that cannot be detected by such graphical criteria, consider adding to D_p the rule $p \wedge b \to f$. Again no rule will be tolerated and the set will be proclaimed p-inconsistent, thus showing (by Thm. 2.8) that $p \wedge b \to \neg f$ is p-entailed by D_p as expected ("typically penguin-birds don't fly"). Interestingly, all these conclusions remain valid upon changing Rule 2 into a strict conditional $p \Rightarrow b$ (which is the usual way of representing the penguin triangle), showing that strict class subsumption is not really necessary for facilitating specificity-based preferences in this example.

Example 2.3 (Quakers and Republicans) Consider the following set of rules:

1. $n \to r$ ("typically Nixonites[7] are Republicans").

2. $n \to q$ ("typically Nixonites are Quakers").

3. $q \Rightarrow p$ ("all Quakers are pacifists").

4. $r \Rightarrow \neg p$ ("all Republicans are nonpacifists").

5. $p \to c$ ("typically pacifists are persecuted").

Rule 5 is tolerated by all others, but the remaining rules are not confirmable, hence inconsistent. The following modification renders the knowledge base consistent:

1. $n \Rightarrow r$ ("all Nixonites are Republicans").

2. $n \Rightarrow q$ ("all Nixonites are Quakers").

3. $q \to p$ ("typically Quakers are pacifists").

4. $r \to \neg p$ ("typically Republicans are nonpacifists").

---

[7]"Nixonites" is shorthand for people who share aspects of Richard M. Nixon's cultural background.

5. $p \rightarrow c$ ("typically pacifists are persecuted").

Indeed, there is a basic conceptual difference between the former case and this one. If all Quakers are pacifists and all Republicans are nonpacifists, our intuition immediately reacts against the idea of finding an individual who is both a Quaker and a Republican. The modified knowledge base, on the other hand, allows a Nixonite who is both a Quaker and a Republican to be either pacifist or nonpacifist. Note that both $n \rightarrow p$ and $n \rightarrow \neg p$ are consistent when added to the knowledge base, so neither one is p-entailed and we can assert that the conclusion is *ambiguous* (i.e., we cannot decide whether a Nixonite is typically a pacifist or not).

Finally, if we make Rules 2 and 4 the only strict rules, we get a knowledge base similar in *structure* to the example depicted by network Γ_6 in [58]:

1. $n \rightarrow r$ ("typically Nixonites are Republicans").

2. $n \Rightarrow q$ ("all Nixonites are Quakers").

3. $q \rightarrow p$ ("typically Quakers are pacifists").

4. $r \Rightarrow \neg p$ ("all Republicans are nonpacifists").

5. $p \rightarrow c$ ("typically pacifists are persecuted").

Not surprisingly, the criterion of Theorem 2.4 renders this knowledge base consistent and $n \rightarrow \neg p$ is p-entailed in conformity with the intuition expressed in [58].

2.6 Reasoning with p-Inconsistent Knowledge Bases

The theory developed in previous sections presents desirable features from both the semantic and computational standpoints. However, the entailment procedure insists on starting with a p-consistent set of conditional rules. In this section, we relax this requirement and explore two proposals for making entailment insensitive to contradictory statements in unrelated portions of the knowledge base, so that mistakes in the encoding of properties about penguins and birds would not tamper with our ability to reason about politicians (e.g., Quakers and Republicans). The first proposal amounts to accepting local p-inconsistencies as deliberate albeit strange expressions, while the second treats them as programming "bugs".

In Def. 2.1 a conditional rule $\varphi \rightarrow \psi$ was assigned the conditional probability $P(\psi|\varphi)$ if P was *proper* for $\varphi \rightarrow \psi$ (i.e., if $P(\varphi) > 0$). In our first proposal for

reasoning with p-inconsistent knowledge bases, we will regard improper probability assignments as admissible and define $P(\psi|\varphi) = 1$ whenever $P(\varphi) = 0$.[8] With this approach, any set Δ, as long as $\hat{\Delta}$ is logically satisfiable,[9] can be represented by the trivial, high probability distribution in which some antecedents receive zero probability. Also, strict rules such as $\phi \Rightarrow \sigma$ can be represented as $\phi \wedge \neg\sigma \rightarrow False$, since we can now use $P(\phi \wedge \neg\sigma) = 0$ to get $P(\sigma|\phi) = 1$. As before, we say that a rule $\varphi \rightarrow \psi$ is *implied*[10] by a (possibly p-inconsistent) set Δ if $\varphi \rightarrow \psi$ receives arbitrarily high probability in all probability assignments in which rules in Δ receive arbitrarily high probability.

Definition 2.15 (p_1-Implication) Given a set Δ and a rule $\varphi' \rightarrow \psi'$, Δ p_1-*implies* $\varphi' \rightarrow \psi'$, written $\Delta \models_{p_1} \varphi' \rightarrow \psi'$, if for all $\varepsilon > 0$ there exists a $\delta > 0$ such that for all probability assignments P, if $P(\psi|\varphi) \geq 1 - \delta$ for all $\varphi \rightarrow \psi \in \Delta$ and $P(\sigma|\phi) = 1$ for all $\phi \Rightarrow \sigma \in \Delta$, then $P'(\psi'|\varphi') \geq 1 - \varepsilon$.
□

The only difference between Def. 2.15 and that of p-entailment (Def. 2.7) is that none of the probability assignments in the definition above are constrained to be proper.

Any p-inconsistent Δ will have a nonempty subset violating one of the conditions of Theorem 2.4. Given that almost all properties stated in this section will refer to such sets, we find it convenient to introduce the following definition:

Definition 2.16 (Unconfirmable set) $\Delta = D \cup S$ is said to be *unconfirmable* if one of the following conditions is true:

1. If D is nonempty, then there cannot be a defeasible rule in D that is tolerated by Δ.

2. If D is empty (i.e., $\Delta = S$), then there must be a strict rule in S that is not tolerated by Δ.

□

---

[8]Even though $P(\varphi \rightarrow \psi) = 1$ if $P(\varphi) = 0$ in Def. 2.1, $P(\varphi \rightarrow \psi)$ was not related to a conditional probability in those cases.

[9]If $\hat{\Delta}$ is not satisfiable this proposal cannot do better than propositional logic, that is, any conditional rule will be trivially entailed.

[10]We will use the term "implication" instead of "entailment" to stress the fact that the set of premises may constitute a p-inconsistent set. For simplicity, however, we will keep the symbol $\models$.

Note that a set Δ_u can be unconfirmable, while both a superset of Δ_u or one of its subsets can be confirmable. The problem of deciding whether a rule is p_1-implied is no worse than that of deciding p-entailment, as shown by the next theorem (proven in [48]).

Theorem 2.17 Δ p_1*-implies* $\varphi \to \psi$ *iff* $\varphi \to \neg\psi$ *belongs to an unconfirmable subset of* $\Delta \cup \{\varphi \to \neg\psi\}$.

This unconfirmable subset can be identified using the p-consistency test discussed in Section 2.4 (Fig. 2.1), and it follows that p_1-implication also requires a polynomial number of satisfiability tests. Moreover, p-entailment is equivalent to p_1-implication if Δ is p-consistent (see Thm. 2.23). For example, consider the union of $D_p = \{b \to f, p \to b, p \to \neg f\}$ of Example 2.2 (encoding the penguin triangle) and the p-inconsistent set $D_i = \{\phi \to \psi, \phi \to \neg\psi, \phi \to \sigma\}$. Some of the rules p_1-implied by $\Delta_i = D_p \cup D_i$ are $p \wedge b \to \neg f$ ("typically, penguin-birds don't fly"), $b \to \neg p$ ("typically birds are not penguins"), and $\phi \to \sigma$. Some of the rules *not* p_1-implied by Δ_i are $p \wedge b \to f$ and $p \to \psi$. Thus, despite its p-inconsistency, not all rules are p_1-implied by Δ_i. However, this example also demonstrates a disturbing feature of p_1-implication: Not only $\phi \to \psi$ and $\phi \to \neg\psi$ but also $\phi \to \neg\sigma$ and $\phi \to p$ (where p is any predicate) are p_1-implied. Thus, although the natural properties of penguins remain unperturbed by the p-inconsistency of D_i, strange rules such as $\phi \to p$ are deduced even though there is no argument to support them (see [58] for similar considerations on inconsistent rules in the context of inheritance networks).

To locate the source of this phenomenon, it is useful to declare a formula to be *inconsistent* if the formula is *False* by default.

Definition 2.18 (Inconsistent formula) Given a set Δ and a formula ϕ, we say that ϕ is an *inconsistent formula* with respect to Δ iff Δ p_1-implies $\phi \to False$.
□

The next theorem relates p_1-implication to Definition. 2.18 and provides an alternative definition of inconsistent formulas in terms of propositional entailment. It is an easy consequence of Theorem 2.17.

Theorem 2.19 *Consider a set* Δ *of conditional rules and the formulas* σ *and* ψ:

1. $\Delta \models_{p_1} \sigma \to \psi$ *iff* σ *is an inconsistent formula with respect to* $\Delta \cup \{\sigma \to \neg\psi\}$.

2. *If* σ *is an inconsistent formula with respect to* Δ, *any conditional rule with* σ *as antecedent will be* p_1*-implied by* Δ.

3. A formula σ is inconsistent with respect to a set Δ iff there exists an unconfirmable subset Δ' of Δ such that $\hat{\Delta'} \models \neg\phi$ [11] where σ is the antecedent of a rule in Δ'.

Theorem 2.19.2 explains why a rule such as $\phi \to p$ is p_1-implied by Δ_i: ϕ is an inconsistent formula with respect to Δ_i, hence *any* rule with ϕ as antecedent will be trivially p_1-implied by Δ_i.

This deficiency of p_1-implication is removed in p_2-*implication*, our second proposal for reasoning with p-inconsistent knowledge bases. The intuition behind p_2-implication is that a rule is considered "implied" only if its negation would introduce a *new* p-inconsistency into the knowledge base. Previous p-inconsistencies are thus considered as programming glitches and are simply ignored.

Definition 2.20 (p_2-Implication) Given a set Δ, we say that $\phi \to \psi$ is p_2-*implied* by Δ, written $\Delta \models_{p_2} \phi \to \psi$, iff $\phi \to \psi$ is not an inconsistent rule with respect to Δ (see Def 2.14) but its negation $\phi \to \neg\psi$ is.
$\square$

The requirement that not both $\phi \to \psi$ and $\phi \to \neg\psi$ be inconsistent serves two purposes. First, as with p-entailment, it constitutes a *safeguard* against rules being trivially implied by virtue of their antecedents being false. Second, if both rules are inconsistent, the contradiction that originates when either is added to Δ must previously have been embedded in Δ and therefore cannot be *new*. In our previous example, the rules $p \wedge b \to \neg f$, $b \to \neg p$, and $\phi \to \sigma$ are p_2-implied by Δ_i; however, contrary to p_1-implication, the rules $\phi \to \psi$, $\phi \to \neg\psi$, $\phi \to \neg\sigma$, and $\phi \to p$ are not. As stated in Theorem 2.23, p_2-implication is strictly stronger than p_1-implication and is equivalent to p-entailment if the set Δ is p-consistent.

Since the notion of p_2-implication is based on the concept of an inconsistent rule (Def. 2.14), there is strong evidence that any procedure for deciding p_2-implication will be exponential (see Sec. 2.4). To obtain a more efficient decision procedure, we propose to weaken the definition of an inconsistent rule. Instead of testing whether a given rule is responsible for a p-inconsistency, we will test whether the rule is responsible for creating an inconsistent formula (see Thm. 2.19).

Definition 2.21 (Weakly inconsistent rule) The rule r is *weakly inconsistent* with respect to a set Δ, iff there exists an unconfirmable subset Δ_u of

---

[11] Recall that $\hat{\Delta}$ denotes the conjunction of the material counterparts of the conditional rules in Δ.

$\Delta \cup \{r\}$, such that $\hat{\Delta}_u \models \neg\phi$ but $\hat{\Delta}'_u \not\models \neg\phi$, where $\Delta'_u = \Delta_u - \{r\}$ and ϕ is the antecedent of some rule in Δ_u.
□

This leads naturally to the notion of weak p_2-implication.

Definition 2.22 (wp_2-Implication) Given a set Δ, a rule $\phi \rightarrow \psi$ is wp_2-*implied* by Δ, written $\Delta \models_{wp_2} \phi \rightarrow \psi$, iff $\phi \rightarrow \neg\psi$ is weakly inconsistent with respect to Δ.
□

As in both p_1- and p_2-implication, the set $\Delta_i = D_p \cup D_i$ wp_2-implies the rules $p \wedge b \rightarrow \neg f$, $b \rightarrow \neg p$, and $\phi \rightarrow \sigma$. More importantly, contrary to p_1-implication (but similar to p_2-implication), the undesirable rules $\phi \rightarrow \neg\sigma$ and $\phi \rightarrow p$ are not wp_2-implied by Δ_i and, in general, wp_2-implication will not sanction a rule merely because its antecedent is inconsistent. However, unlike p_2-implication, wp_2-implication will sanction any rule whose consequence is the negation of an inconsistent formula (for example, $p \rightarrow \neg\phi$).

The notion of wp_2-implication is situated somewhere between p_1-implication and p_2-implication, as the next two theorems indicate. It rests semantically on both, since it requires the concepts of inconsistent formulas and inconsistent rules. It also preserves some of the computational advantages of p_1-implication.

Theorem 2.23   *1. Given a p-consistent set Δ, the notions of p-entailment, p_1-implication, wp_2-implication, and p_2-implication are equivalent.*

2. Given a p-inconsistent set Δ, p_2-implication is strictly stronger than wp_2-implication, and wp_2-implication is strictly stronger than p_1-implication.

Theorem 2.24 *If the set Δ is acyclic and of Horn form, wp_2-implication can be decided in polynomial time.*

The need to search for a suitable unconfirmable subset Δ_u (see Def. 2.22) results in wp_2-implication being computationally harder than p_1-implication.

2.7   Discussion

We have formalized a norm of consistency for mixed sets of conditionals, ensuring that every group of rules is satisfiable in a non-trivial way, one in which the antecedent and the consequent of at least one rule are both true. We showed that

any group of rules not satisfiable this way must contain conflicts that cannot be reconciled by appealing to exceptions or ambiguities and is thus normally considered contradictory (i.e., unfit to represent world knowledge). Using this norm, we devised an effective procedure to test for inconsistencies and established a tight relation between entailment and consistency, permitting entailment to be decided by using consistency tests. These tests were shown to require polynomial complexity relative to propositional satisfiability. We also discussed ways of drawing conclusions from inconsistent knowledge bases and of uncovering sets of rules directly responsible for such inconsistencies.

One of the key requirements in our definition of consistency is that no conditional rule in Δ should have an impossible antecedent and, moreover, that no antecedent should become absolutely impossible as exceptions (to defeasible rules) become less likely (i.e., as ε becomes smaller). This requirement reflects our understanding that it is fruitless to build knowledge bases for nonexisting classes and counterintuitive to deduce (even defeasibly) conditional rules having impossible antecedents. Consequently, pairs such as $\{\phi \rightarrow \psi, \phi \rightarrow \neg\psi\}$ or $\{\phi \Rightarrow \psi, \phi \Rightarrow \neg\psi\}$ are labeled inconsistent and treated as unintentional mistakes. The main application of the procedures proposed in this chapter is to alert users and knowledge providers of such glitches, in order to prevent undesirable inferences.

This chapter also presents a new formalization for strict conditional rules, within the analysis of probabilistic consistency, that is totally distinct from their material counterparts. The importance of this distinction has been recognized by several researchers (see [104, 23, 35, 37] and others) and has both theoretical and practical implications.

In ordinary discourse, conditionals are recognized by universally quantified subsumptions such as "all penguins are birds" or, in the case of ground rules, by the use of the English word "if" (e.g., "if Tweety is a penguin, then she is a bird"). The function of these *indicators* is to alert the listener that the assertion made is not based on evidence pertaining to the specific individual, but rather on generic background knowledge pertaining to the individual's class (e.g., being a penguin). It is this pointer to the background information that is lost if a conditional rule is encoded as a Boolean expression, and it is this information that is crucial for adequately processing specificity preferences.

Intuitively, background knowledge encodes the general tendency of things to happen (i.e., relations that hold true in all worlds) while evidential knowledge describes that which actually happened (i.e., relations in our particular world). Thus, conditional rules, both defeasible and indefeasible, play a role similar to

that of meta-inference rules: They tell us how to draw conclusions from specific observations about a particular situation or a particular individual, but do not themselves convey such observations. It is for this reason that we chose to use a separate connective, $\Rightarrow$, to denote strict conditionals, as is done in [58] in the context of inheritance networks. Strict conditionals, by virtue of pointing to generic background knowledge, are treated as part of the knowledge base, while propositional formulas, including material implications, are used to formulate queries but are excluded from the knowledge base itself. As a result, the rule $p \Rightarrow b$ is treated as a constraint over the set of admissible probability assignments, while the propositional formula $p \supset b$ is treated as specific evidence or a specific observation on which these probability assignments are to be conditioned.

It does indeed make a profound difference whether our knowledge of *Tweety*'s birdness comes from generic background knowledge about penguins or from specific observations conducted on *Tweety*. In natural language, the latter case would normally be phrased by nonconditional rules such as "it is not true that *Tweety* is both a penguin and a non-bird", which is equivalent to the material implication $penguin(Tweety) \supset bird(Tweety)$.

The practical aspects of this distinction can best be demonstrated using the penguin example (Ex. 2.2).[12] Assume we know that "typically birds fly" and "typically penguins do not fly". If we are told "Tweety is a penguin" and "all penguins are birds", we would like to conclude that Tweety does not fly. By the same token, if we are told "Tweety is a bird" and "all birds are penguins", we would have to conclude that Tweety does fly. However, note that both $\{p, p \supset b\}$ and $\{b, b \supset p\}$ are logically equivalent to $\{p, b\}$, which totally ignores the relation between penguins and birds and should yield identical conclusions regardless of whether penguins are subclass of birds or the other way around. Thus, when treated as material implications, information about class subsumption is permitted to combine with properties attributed to individuals and therefore this crucial information gets lost.

This distinction was encoded in [37] by placing strict conditionals together with defaults in a *background context*, separate from the *evidential set* which was reserved for observations made on a particular state of affairs. In [72, p. 212] it is stated that "dealing with hard constraints, in addition to soft ones, involves relativizing to some given set of tautologies". Here, again, strict conditionals and ground formulas would receive different treatment; only the former are permitted to influence rankings among worlds. The separate connective $\Rightarrow$ used in this chapter treatment makes this distinction clear and natural, and the uniform prob-

---

[12]Taken from [37].

abilistic semantics given to both strict and defeasible rules adequately captures the notion of consistency in systems containing such mixtures.

In Chapter 5, the distinction between $\supset$ and $\Rightarrow$ is crucial since rules are used to impose a Markovian condition of independence among the atomic propositions in the language, in order to induce causal interpretations, while wffs are used to provide the *context* of a query.

The notion of p-entailment is known to yield a rather conservative set of conclusions. For example, let $\Delta' = \{a \rightarrow b\}$, and let a, b, c be atomic propositions in $\mathcal{L}$. It seems reasonable to expect $\Delta' \models_p c \wedge a \rightarrow b$ simply because c represents an *irrelevant* proposition, one with no relation to our knowledge base Δ'. Yet, the rule $c \wedge a \rightarrow b$ is not p-entailed by Δ. The reason is that the notion of p-entailment requires $P(b|a \wedge c)$ to attain arbitrarily high probability in all those probability distributions in which $P(b|a)$ attains arbitrarily high probability. A probability distribution P' where $P'(b|a) \geq 1 - \varepsilon$ (for every $\varepsilon > 0$) and yet $P'(b|a \wedge c) = 0$ can be easily built.[13] For this reason, p-entailment is not proposed as a complete characterization of defeasible reasoning. It nevertheless yields a core of plausible consequences that should be maintained in every system that reasons defeasibly [97, 98]. Similar problems with irrelevance are shared by all other proposals for nonmonotonic reasoning based on a conditional interpretation of the rules (see, for example, [23, 69, 14, 36]). Extensions of p-entailment presenting solutions to these problems will be explored in Chapters 3, 4, and 5.[14] Nonprobabilistic extensions can be found in [72, 36, 14]. All these formalisms, as well as circumscription [88], default logic [108], and argument-based systems [84, 58], could benefit from a preliminary test of consistency such as the one proposed in this chapter.

---

[13]This is not surprising since $a \rightarrow b$ does not say much about the worlds for c or $\neg c$.

[14]Chapters 3 and 4 are independent of each other and can be read in any order.

CHAPTER 3

Plausibility I: A Maximum Entropy Approach

3.1   Introduction

This chapter proposes an approach to nonmonotonic reasoning that combines the principle of infinitesimal probabilities (described in Chap. 2) with the principle of maximum entropy in order to extend the inferential power of the probabilistic interpretation of defaults. As pointed out in Sections 1.1 and 1.2, conditional based approaches (such as p-entailment) fail to sanction some desirable patterns which are readily sanctioned in common discourse. Their main weakness stems from the failure to properly handle irrelevant information (see Sec. 2.7 and [101, 38]). Recent extensions (Delgrande's proposal [23], rational closure [72], and system-Z [100]) to the conditional based approaches were successful in capturing some aspects of irrelevance, but still suffer from the *ills* of conservativism, namely, they fail to sanction *property inheritance* from classes to exceptional subclasses. For example, given that "birds fly", "birds have beaks", and "penguin-birds don't fly", these extensions fail to conclude that penguin-birds have beaks (despite their being exceptional relative to flying). The maximum entropy formalism described in this chapter is proposed as a well-disciplined approach to extracting implicit (probabilistic based) independencies from fragments of knowledge, so as to overcome those ills. In this respect, the resulting formalism combines the virtues of both the extensional and the conditional based approaches (see Sec. 1.2 for a review of both approaches to default reasoning).

The connection between maximizing entropy and minimizing dependencies has been recognized by several workers [63, 123] and was proposed for default reasoning by Pearl [97, p. 491]. The origin of this connection lies in statistical mechanics, where the entropy approximates the (logarithm of) the number of distinct configurations (assignments of properties to individuals) that comply with certain constraints [12]. For example, if we observe that in a certain population the proportion of tall individuals is p and the proportion of smart individuals is q, then out of all configurations that comply with these observations, those in which the proportion of smart-and-tall individuals is pq (as dictated by the assumption of independence) constitute the greatest majority; any other proportion of smart-

and-tall individuals would permit the realization of fewer configurations, and will correspond to a lower entropy value.

In physics, a configuration stands for the assignment of each particle to a particular cell in position-momentum space, and all distinguishable configurations can be assumed to have equal a priori probabilities. Therefore, the maximum entropy distribution also represents the *most likely* distribution, that is, the distribution most likely to be found in nature. Indeed, the celebrated distributions of Maxwell-Boltzman and Fermi-Dirac, both maximizing the entropy under the appropriate assumptions, have been observed to hold with remarkable accuracy and stability.

A similar argument can be invoked to justify the use of maximum entropy in reasoning applications where possible worlds play the role of configurations, and the constraints are given by observed statistical proportions, for example, "90% of all birds fly" (see Bacchus et al. [7]). Likewise, if we assume that default expressions are qualitative abstractions of probabilities, and that probabilities are degrees of belief that reflect proportions in an agent's experience, then it is also reasonable to assume that our interpretation of defaults, as manifested in discourse conventions, is governed by principles similar to that of maximum entropy. In view of the "most likely" status of the maximum entropy distribution, it is quite possible that discourse conventions have evolved to conform with the maximum entropy principle for pragmatic reasons; conformity with this principles assures conformity with the highest number of experiences consistent with the available defaults.

The chapter is organized as follows: Section 3.2 recasts the notions behind p-entailment in terms of consequence relations and parameterized probability distributions (PPD).[1] This reformulation has the advantages of conceptual simplicity and expressiveness. Each PPD will induce a consequence relation $\phi \mathrel{\vert\!\sim} \sigma$ on wffs. A query such as "is σ plausible in the context of ϕ, given the knowledge base Δ" will be then evaluated in terms of the set of consequence relations induced by the PPDs *admissible* with the constraints in Δ. In this manner, each probability model for a given knowledge base can be characterized and compared in terms of the plausible conclusions it sanctions. By the same token, comparisons with other formalisms are also facilitated. It is shown that this reformulation preserves all the properties of p-entailment (see, for example, Thm. 3.10), and that consequence relations are enhanced with a desirable property called Rational

---

[1]This special set of distributions present a smoothness property necessary for the computation of the maximum entropy distribution (see Def. 3.1).

Monotony.[2] Section 3.3 is concerned with the symbolic machinery necessary to compute the consequence relation associated with the maximum entropy distribution, and develops such machinery for a class of default rules called *minimal-core* sets. Section 3.4 gives examples illustrating the behavior of the consequence relations that results from maximizing entropy. Section 3.5 discusses issues related to non-minimal-core sets, and Section 3.6 summarizes the main results. A step-by-step application of the Lagrange multipliers technique can be found in Appendix B, while Appendix A contains proofs of the main theorems and propositions.

3.2 Parameterized Probability Distributions

The definition below restricts the set of *acceptable* probability distributions to those that present an analytic property around $\varepsilon = 0$. This restriction is convenient for computing maximum entropy and for introducing the concepts of consequence relations (see Def. 3.4) and rankings (see Def. 3.7). As is demonstrated in Theorems 3.3 and 3.10, this restriction does not affect the notions of p-consistency and p-entailment introduced in Chapter 2.

Definition 3.1 (Parameterized probability distribuion) A *parameterized probability distribution* (PPD) is a collection $\{P_\varepsilon\}$ of probability measures over the set Ω of possible worlds, indexed by a parameter ε. $\{P_\varepsilon\}$ assigns to each possible world ω a function of ε, $P_\varepsilon(\omega)$, such that:

1. $P_\varepsilon(\omega) \geq 0$ for all $\varepsilon > 0$, and

$$\sum_{\omega \in \Omega} P_\varepsilon(\omega) = 1 \quad \text{for all } \varepsilon > 0. \tag{3.1}$$

2. For every ω, $P_\varepsilon(\omega)$ is analytic at $\varepsilon = 0$. In other words, PPD's can be expanded as a Taylor series about zero.

□

For each formula $\varphi \in \mathcal{L}$, $P_\varepsilon(\varphi)$ is defined as

$$P_\varepsilon(\varphi) = \sum_{\omega \models \varphi} P_\varepsilon(\omega), \tag{3.2}$$

---

[2]Ginsberg [44] however, argues against Rational Monotony.

and for each ψ and φ in $\mathcal{L}$, the conditional probability $P_\varepsilon(\psi|\varphi)$ is defined as

$$P_\varepsilon(\psi|\varphi) = \begin{cases} 1 & \text{if } P_\varepsilon(\varphi) = 0 \\ \frac{P_\varepsilon(\psi \wedge \varphi)}{P_\varepsilon(\varphi)} & \text{otherwise} \end{cases} \tag{3.3}$$

For simplicity P_ε will be used when referring to the PPD (strictly, $\{P_\varepsilon\}$) in the remainder of the chapter, wherever the distinction is clear from the context.

Definition 3.2 (Consistency) A given P_ε is *admissible* with respect to a set D iff for each $\varphi_i \to \psi_i \in D$,

$$\lim_{\varepsilon \to 0} P_\varepsilon(\psi_i|\varphi_i) = 1 \tag{3.4}$$

and $P_\varepsilon(\varphi_i) > 0$. D is said to be *consistent* iff there exists at least one admissible P_ε with respect to D.[3]
□

Theorem 3.3 *A set D is* consistent *iff D is* p-consistent.

It follows that procedure Test_Consistency outlined in Figure 2.1 can be used to check consistency as defined in Definition 3.2. The definitions above provide a semantical interpretation for each default rule in D in terms of infinitesimal conditional probabilities, where ψ accrues an arbitrarily high likelihood whenever φ is all that is known.

As Theorem 3.10 will show, Defs. 3.1 and 3.4 recast the notions of p-entailment in terms of consequence relations. The study of nonmonotonic and default reasoning in terms of consequence relations was first proposed by Gabbay [32] and further explored in [85, 69, 72].

Definition 3.4 (Consequence relations) Every P_ε induces a unique *consequence relation* $\mathrel{\vdash\!\!\!\sim}$ on formulas, defined as follows:

$$\phi \mathrel{\vdash\!\!\!\sim} \sigma \quad \text{iff} \quad \lim_{\varepsilon \to 0} P_\varepsilon(\sigma|\phi) = 1 \tag{3.5}$$

$\varphi \mathrel{\vdash\!\!\!\sim} \psi$ is *proper* in P_ε if $P_\varepsilon(\varphi) > 0$ for all ε. The set of proper $\varphi \mathrel{\vdash\!\!\!\sim} \psi$ will be called the proper consequence relation of P_ε.
□

---

[3]Note that we are only dealing with defeasible rules. The generalization to strict rules follows immediately from the concepts introduced in Chapter 2. We just need to augment the conditions for admissibility in Def. 3.2 to include the requirement that for each $\phi_j \Rightarrow \sigma_j \in S$, $P_\varepsilon(\sigma_j|\phi_j) = 1$.

Note that a given P_ε is *admissible* with respect to a set D iff for each $\varphi_i \to \psi_i \in D$, $\varphi_i \mathrel{\vdash\!\!\!\sim} \psi_i$ belongs to the proper consequence relation of P_ε.

Among the general *rules of inference* that a commonsense consequence relation might be expected to obey, the following have been proposed [37, 69, 85]:[4]

(Logic) If $\varphi \vdash \psi$, then $\varphi \mathrel{\vdash\!\!\!\sim} \psi$.

(Cumulativity) If $\varphi \mathrel{\vdash\!\!\!\sim} \psi$, then $\varphi \mathrel{\vdash\!\!\!\sim} \gamma$ iff $\varphi \wedge \psi \mathrel{\vdash\!\!\!\sim} \gamma$.

(Cases) If $\varphi \mathrel{\vdash\!\!\!\sim} \psi$ and $\gamma \mathrel{\vdash\!\!\!\sim} \psi$, then $\varphi \vee \gamma \mathrel{\vdash\!\!\!\sim} \psi$.

Kraus et al. [69] introduce a class of preferential models and show that each preferential model satisfies the three laws given above. Moreover, they show that every consequence relation satisfying those laws can be represented as a preferential model.[5] Analogous results were shown in [72] with respect to the class of ranked preferential models and the set of rules above augmented by the rule of rational monotony:

$$\text{If } \varphi \mathrel{\vdash\!\!\!\sim} \psi \text{ and } \varphi \mathrel{\not\vdash\!\!\!\sim} \neg\gamma, \text{ then } \varphi \wedge \gamma \mathrel{\vdash\!\!\!\sim} \psi. \tag{3.6}$$

Theorems 3.5 and 3.6 formalize the relation between a consequence relation induced by a PPD and a consequence relation satisfying the rules of inference above. In particular, Theorem 3.6, which follows from the results in [74], constitutes a representation theorem stating that any finite consequence relation that satisfies logic, cumulativity, cases, and rational monotony can be represented by a PPD.[6]

Theorem 3.5 *A PPD consequence relation satisfies the* logic, cumulativity, cases, and rational monotony *rules of inference.*

Theorem 3.6 *Every PPD consequence relation can be represented as a ranked preferential model, and every ranked preferential model with a finite non-empty state space can be represented as a PPD consequence relation.*

We remark that ε-semantics, as defined in [37, 97] does not comply in general with the rule of rational monotony. This is because ε-entailment was defined as the intersection of the consequence relations induced by *all* admissible P_ε,

---

[4]Geffner [37] proposes an additional rule of inference: If $\varphi \to \psi \in D$, then $\varphi \mathrel{\vdash\!\!\!\sim} \psi$. This rule establishes a connection between the defaults in the knowledge base and the consequence relation $\mathrel{\vdash\!\!\!\sim}$. Semantically, this connection is established in this dissertation by interpretating defaults in D as constraints over rankings (see Defs. 3.2 and 3.4).

[5]They actually use a slightly different set of inference rules which can be shown to be equivalent to those above.

[6]Similar results have been obtained independently by Satoh [113].

and P_ε was not restricted to analytic functions. Thus, the set of admissible P_ε's included discontinuous functions, for which the $\lim_{\varepsilon \to 0}$ in Eq. 3.5 does not exist. By restricting the probability distributions to be analytic at $\varepsilon = 0$, we can guarantee that each one of their consequence relations satisfies the rule of rational monotony.

As ε approaches zero, the Taylor series expansion of P_ε is dominated by the first term whose coefficient is non-zero. Thus, we can define a ranking function on possible worlds using the exponent of this dominant term as follows.

Definition 3.7 (Ranking) Given a P_ε, the ranking function $\kappa_{P_\varepsilon}(\omega)$ is defined as

$$
\kappa_{P_\varepsilon}(\omega) = \begin{cases} \min\{n \text{ such that } \lim_{\varepsilon \to 0} \frac{P_\varepsilon(\omega)}{\varepsilon^n} \neq 0\} & \text{if } P_\varepsilon(\omega) > 0 \\ \infty & \text{if } P_\varepsilon(\omega) = 0 \end{cases}
\tag{3.7}
$$

$\square$

Moreover, according to Eq. 3.2, P_ε also induces a ranking on formulas φ:

$$
\kappa_{P_\varepsilon}(\varphi) = \min_{\omega \models \varphi} \kappa_{P_\varepsilon}(\omega)
\tag{3.8}
$$

Proposition 3.8 *The following are consequences of Defs. 3.1, 3.4, and 3.7. Given P_ε:*

1. *There is at least one possible world ω such that $\kappa_{P_\varepsilon}(\omega) = 0$.*

2. *$\phi \hspace{1pt}\vdash\hspace{-6pt}\sim\hspace{1pt} \sigma$ holds in P_ε iff either $\kappa_{P_\varepsilon}(\phi \wedge \sigma) < \kappa_{P_\varepsilon}(\phi \wedge \neg\sigma)$ or $\kappa_{P_\varepsilon}(\phi) = \infty$.*

3. *We will say that κ_{P_ε} is admissible with respect to D if for each $\varphi_i \to \psi_i \in D$*

$$
\kappa_{P_\varepsilon}(\varphi_i \wedge \psi_i) < \kappa_{P_\varepsilon}(\varphi_i \wedge \neg\psi_i).
\tag{3.9}
$$

   D is consistent iff there exists at least one admissible ranking κ_{P_ε} with respect to D.

Expressed in terms of ranking functions, the consequence relation induced by a PPD echoes the preferential interpretation for defeasible sentences advocated in [117] according to which ψ should hold in all minimal (preferred, more normal) models for φ. This can be seen more clearly by writing Eq. 3.9 as $\kappa(\varphi) < \kappa(\varphi \wedge \neg\psi)$ and recalling that, in our case, minimality (preference or normality) is reflected in having the lowest possible ranking (i.e., the highest possible likelihood).

The rankings induced by P_ε will prove useful in the maximum entropy computation process. Rather than computing the PPD of maximum entropy, P^*, we will calculate its corresponding ranking κ_{P^*} (denoted κ^*), from which we can compute the consequence relation associated with P^*.

3.3 Plausible Conclusions and Maximum Entropy

Given that ϕ is true, a formula σ will be *probabilistically entailed* by D if

$$\lim_{\varepsilon \to 0} P_\varepsilon(\sigma|\phi) = 1$$

in all P_ε that are admissible with respect to D.

Definition 3.9 (Probabilistic entailment) Given a consistent D, a formula σ is *probabilistically entailed* by D given ϕ, written $\phi \mathrel{|\!\!\sim_p} \sigma$, iff $\phi \mathrel{|\!\!\sim} \sigma$ is in the proper consequence relation of *all* P_ε admissible with D.
$\square$

As is expected there is a close relation between p-entailment (Def. 2.7 and probabilistic entailment:

Theorem 3.10 *Given a consistent D, $\phi \mathrel{|\!\!\sim_p} \sigma$ iff $D \models_p \phi \to \sigma$.*

It follows that we can use the decision procedure for p-entailment (based on procedure **Test_Consistency** in Figure 2.1) for deciding probabilistic entailment.[7] Probabilistic entailment yields (semimonotonically) the most conservative "core" of plausible conclusions that one would wish to draw from a conditional database if one is committed to avoiding inconsistencies [101]. In particular, it does not permit chaining (from $a \to b$ and $b \to c$ conclude $a \mathrel{|\!\!\sim} c$), or contraposition (from $a \to b$ conclude $\neg b \mathrel{|\!\!\sim} \neg a$), hence it is too weak to be proposed as a complete characterization of defeasible reasoning. As was mentioned in Sections 1.1 and 2.7, the reason for this conservative behavior lies in our insistence that any conclusion must attain high probability in *all* probability distributions admissible with D. Thus, given $D = \{p \to \neg f\}$ ("typically penguins don't fly") and the proposition bl (for "blue"), the conclusion $bl \wedge p \mathrel{|\!\!\sim} \neg f$ ("blue penguins don't fly") will not be sanctioned by probabilistic entailment, since one admissible distribution reflects a *world* in which blue penguins do fly. Clearly, if we want the system to

---

[7]This notion of entailment is also equivalent to the notion of preferential entailment in [69], even though preferential entailment was motivated by considering desirable features of the consequence relations and not by probabilistic considerations. The relation between these two notions was reported by Kraus et al. [69].

respect the communication convention that, unless stated explicitly, properties are presumed to be *irrelevant* to each other, we need to restrict the family of probability distributions relative to which a given query is evaluated. In other words, we should consider only distributions that minimize dependencies, that is, they should contain the dependencies that are absolutely implied by D, but no others.

Since the maximum entropy distribution exhibits this minimal commitment to dependencies, it well be the focal point of the inference procedure. The entropy associated with a distribution P_ε is defined as

$$H[P_\varepsilon] = - \sum_\Omega P_\varepsilon(\omega) \log P_\varepsilon(\omega). \tag{3.10}$$

Given a set $D = \{\varphi_i \to \psi_i\}$, the objective is to compute the PPD among those satisfying the constraints imposed by D that maximizes the entropy;[8] P^* denotes this maximum entropy distribution. The formulas in the consequence relation of P^* (denoted by $\mathrel{|\!\sim}$) are then taken as plausible conclusions of D those . In Eq. 3.5, the default $\varphi_i \to \psi_i$ is interpreted as a constraint on the limit of $P_\varepsilon(\psi_i|\phi_i)$ as ε approaches zero. To facilitate the maximization of $H[P_\varepsilon]$, these constraints are replaced by equivalent constraints that assign a specific bound to $P_\varepsilon(\psi_i|\phi_i)$ for every $\varepsilon > 0$. The (unique) maximum entropy distribution for each value of $\varepsilon > 0$ is then derived, and finally, its asymptotic solution as ε approaches zero is examined.

A PPD P_ε, *satisfies* a rule $r_i : \varphi_i \to \psi_i$ iff

$$P_\varepsilon(\psi_i|\varphi_i) \geq \frac{1}{1 + C_i\varepsilon}, \tag{3.11}$$

where C_i is an arbitrary positive coefficient independent of ε. Accordingly, the admissibility constraints (Eq. 3.5) are re-written as:

$$C_i \times \varepsilon \times P_\varepsilon(\psi_i \wedge \varphi_i) \geq P_\varepsilon(\neg\psi_i \wedge \varphi_i). \tag{3.12}$$

As we shall see, the equations governing the ranking approximation will be independent of C_i. $\Omega_{r_i}^+$ denotes the set of possible worlds verifying the rule r_i and $\Omega_{r_i}^-$ denotes the set of possible worlds falsifying the rule r_i, the constraint of Eq. 3.12 can be written

$$\sum_{\omega \in \Omega_{r_i}^-} P_\varepsilon(\omega) - [C_i \times \varepsilon \times \sum_{M \in \Omega_{r_i}^+} P_\varepsilon(\omega)] \leq 0. \tag{3.13}$$

---

[8]The PPD of maximum entropy must also satisfy the normalization constraint $\sum_\Omega P_\varepsilon(\omega) = 1$.

43

The problem is to maximize Eq. 3.10 subject to the constraints given in Eq. 3.13; one constraint for each rule in D. The most powerful method for solving such optimization problems is the Lagrange multipliers technique [5]. This technique associates a factor α with each constraint (rule) and yields a distribution P^* that is expressible as a product of these factors [18]. We will see that, under the infinitesimal approximation (i.e., when ε is close to 0), $P^*(\omega)$ will be proportional to the product of the factors (α) associated only with those sentences falsified in ω.

At the point of maximum entropy, the status of a constraint such as Eq. 3.13 can be one of two types: *active*, when the constraint is satisfied as an equality, and *passive*, when the constraint is satisfied as a strict inequality. Passive constraints do not affect the point of maximum entropy and can be ignored (see [5]). Unfortunately, the task of identifying the set of *active* constraints is in itself a hard combinatorial problem. The analysis will begin by assuming a set of active constraints, then we will provide a characterization of knowledge bases called *minimal-core* sets (Def. 3.11) which are guaranteed to impose only active constraints, and postpone the discussion of inactive constraints till Section 3.5.

An application of the Lagrange multiplier technique on a set of n active constraints yields the following expression for each term $P_\varepsilon(\omega)$ (see Appendix B for a step by step derivation):[9]

$$P_\varepsilon(\omega) = \alpha_0 \times \prod_{r_i \in D_\omega^-} \alpha_{r_i} \times \prod_{r_j \in D_\omega^+} \alpha_{r_j}^{(-C_j\varepsilon)} \tag{3.14}$$

where D_ω^- denotes the set of rules falsified in ω and D_ω^+ denotes the set of rules verified in ω. Motivated by Def. 3.4, we look for an asymptotic solution where each α_{r_i} is proportional to $\varepsilon^{Z(r_i)}$ for some non-negative integer $Z(r_i)$,[10] and thus each term of the form $\alpha_{r_j}^{(-C_j\varepsilon)}$ will tend to one as ε tends to zero. The term α_0 is a normalization constant that will be present in each term of the distribution and thus can be safely ignored. Using P_ε' to denote the unnormalized probability function, and taking the limit as ε goes to zero, Eq. 3.14 yields:

$$P_\varepsilon'(\omega) \approx \begin{cases} 1 & D_\omega^- = \emptyset \\ \\ \prod_{r_i \in D_\omega^-} \alpha_{r_i} & \text{otherwise} \end{cases} \tag{3.15}$$

---

[9]In Eq. 3.14, $\alpha_0 = e^{(\lambda_0 - 1)}$ and $\alpha_{r_k} = e^{\lambda_k}$, where λ_0 and λ_k are the actual Lagrange multipliers.

[10]We take a "bootstrapping" approach: if this assumption yields a solution, then the uniqueness of P^* will justify this assumption. Note that the assumption will be satisfied for analytic functions (see Def. 3.1) which eliminate exponential dependencies on ε.

Thus, the probability of a given possible world ω depends only on the rules that are falsified by that possible world. In other words, any two possible worlds that falsify the same set of rules are (asymptotically) equiprobable. Once the α-factors are computed, we can construct an asymptotic approximation of the desired probability distribution (using the ranking functions of Def. 3.7) and determine the consequence relation $\sim$ of D, according to maximum entropy.

To compute the α-factors we substitute the expression for each $P'_\varepsilon(\omega)$ (Eq. 3.15) in each of the active constraint equations (Eq. 3.13), obtaining:

$$\sum_{m \in \Omega^-_{r_i}} [\prod_{r_k \in D^-_\omega} \alpha_{r_k}] = C_i \varepsilon \times \sum_{M \in \Omega^+_{r_i}} [\prod_{r_j \in D^-_\omega} \alpha_{r_j}] \qquad (3.16)$$

A few observations are in order. First, Eq. 3.16 constitutes a system of n equations (one for each active rule) with n unknowns (the α-factors; one for each active rule). Second, since P_ε is analytic at $\varepsilon = 0$, we can extract the dominant term in the Taylor expansion of each element in the products of Eq. 3.16 and write $\alpha_{r_i} \approx a_i \varepsilon^{Z(r_i)}$ where $Z(r_i)$ is a non-negative integer. Our task reduces to that of computing the Z's.[11] Third, we can replace each summation by its dominant term, namely, the term where ε has the minimal exponent. Thus, taking log on both sides of Eq. 3.16 and writing $\log \alpha_{r_i} \approx \log a_i + Z(r_i) \log \varepsilon \approx Z(r_i) \log \varepsilon$ yields

$$\min_{\Omega^-_{r_i}} [\sum_{r_k \in D^-_\omega} Z(r_k)] = 1 + \min_{\Omega^+_{r_i}} [\sum_{r_j \in D^-_\omega} Z(r_j)] \quad 1 \leq i \leq n, \qquad (3.17)$$

where the minimization is understood to range over all possible worlds in $\Omega^-_{r_i}$ and $\Omega^+_{r_i}$, respectively. Note that the constants C_i do not appear in Eq. 3.17; they interact only with the a_i coefficients which will be adjusted accordingly to match the constraints in Eq. 3.12.

Since rule r_i is falsified in each possible world on the left-hand side of Eq. 3.17, $Z(r_i)$ will appear in each one of the $\sum$-terms inside the *min* operation and can be isolated:

$$Z(r_i) + \min_{\Omega^-_{r_i}} [\sum_{\substack{r_k \in D^-_\omega \\ k \neq i}} Z(r_k)] = 1 + \min_{\Omega^+_{r_i}} [\sum_{r_j \in D^-_\omega} Z(r_j)]. \qquad (3.18)$$

Although Eq. 3.18 offers a significant simplification over Eq. 3.16, it is still not clear how to compute the values for the Z's in the most general case. We now introduce a class of rule sets D for which a simple greedy strategy (see procedure Z^*_order in Figure 3.1) can be used to solve the set of equations above.

---

[11] Each probability term $P'_\varepsilon(\omega)$ is asymptotically determined once the values of the Z's are computed (see Eq. 3.15).

Definition 3.11 (Minimal-Core set) D is a *minimal-core (MC) set* iff for each rule $r_i : \varphi_i \rightarrow \psi_i \in D$, its *negation* $\varphi_i \rightarrow \neg\psi_i$ is tolerated by $D - \{\varphi_i \rightarrow \psi_i\}$. Equivalently, for each rule r_i there is a possible world that falsifies r_i and no other rule in D.
□

For example, the set D_p of Example 2.2 is MC: $\omega_1 \models p \wedge b \wedge f$ falsifies $p \rightarrow \neg f$, $\omega_2 \models \neg p \wedge b \wedge \neg f$ falsifies $b \rightarrow f$, and $\omega_3 \models p \wedge \neg b \wedge \neg f$ falsifies $p \rightarrow b$. On the other hand, changing $p \rightarrow \neg f$ to $p \rightarrow f$ renders D_p no longer MC. Clearly, deciding whether a set D is MC takes $|D|$ satisfiability tests. The MC property excludes sets D that contain redundant rules, that is, rules r that are entailed by the P^* computed for $D - \{r\}$. It follows that all default rules in an MC set are active.

Proposition 3.12 *If D is an MC set, then, for all defaults $r : \varphi \rightarrow \psi \in D$, $\varphi \mathrel{|\!\!\sim} \psi$ is not in the consequence relation induced by $D - \{\varphi \rightarrow \psi\}$.*

The MC property guarantees that, for each rule $r_i \in D$, there is a possible world ω_i in which only that rule is falsified. Thus, since the *min* operation on the left-hand side of Eq. 3.18 ranges over all possible worlds ω in which r_i is falsified, the minimum of such possible worlds is ω_i, and the constraint equations for an MC set can be further simplified to

$$Z(r_i) \;=\; 1 + \min_{\Omega_{r_i}^+}[\sum_{r_j \in D_\omega^-} Z(r_j)] \quad 1 \leq i \leq n. \tag{3.19}$$

Note that since the expression $\sum_{r_j \in D_\omega^-} Z(r_j)$ is equal to the exponent of the most significant ε-term in $P^*(\omega)$, from Definition. 3.7 we have that it is actually equal to $\kappa_{P*}(\omega)$, which we denote by κ^*. Thus, Eq. 3.19 can be rewritten as a pair of coupled equations; the first,

$$\kappa^*(\omega) \;=\; \sum_{r_i \in D_\omega^-} Z^*(r_i), \tag{3.20}$$

assigns a ranking $\kappa^*(\omega)$ to each possible world ω once we know the Z^*-values on the rules. The second,

$$Z^*(r_i) \;=\; \min_{M \models \varphi_i \wedge \psi_i} \kappa^*(\omega) + 1, \tag{3.21}$$

assigns a value $Z^*(r_i)$ to each rule $r_i : \varphi_i \rightarrow \psi_i$ once we know the possible world ranking κ^*. We have reduced the computation of the maximum entropy distribution to finding a Z^* function that is consistent with both Eqs. 3.20 and 3.21. Moreover, given the κ^*-ranking, we can decide entailment $\mathrel{|\!\!\sim}$ by the criterion

$$\varphi \mathrel{|\!\!\sim} \psi \quad \text{iff} \quad \kappa^*(\psi \wedge \varphi) < \kappa^*(\neg\psi \wedge \varphi). \tag{3.22}$$

46

Procedure Z^*_order
Input: A consistent MC set D. **Output:** Z^*-ranking on rules.

1. Let D_0 be the set of rules tolerated by D.

2. For each rule $r_i : \varphi_i \rightarrow \psi_i \in D_0$, set $Z(r_i) = 1$ and $\mathcal{R}Z^+ = D_0$.

3. While $\mathcal{R}Z^+ \neq D$, do:

    (a) Let Ω be the set of possible worlds ω, such that ω falsifies rules only in $\mathcal{R}Z^+$ and verifies at least one rule outside of $\mathcal{R}Z^+$; let $\mathcal{R}Z^-_\omega$ denote the set of rules in $\mathcal{R}Z^+$ falsified by ω.

    (b) For each ω, compute

    $$\kappa(\omega) = \sum_{r_i \in \mathcal{R}Z^-_\omega} Z(r_i). \qquad (3.23)$$

    (c) Let ω^* be the possible world in Ω with minimum κ; for each rule $r_i : \varphi_i \rightarrow \psi_i \notin \mathcal{R}Z^+$ that ω^* verifies, compute

    $$Z(r_i) = \kappa(\omega^*) + 1 \qquad (3.24)$$

    and set $\mathcal{R}Z^+ = \mathcal{R}Z^+ \cup \{r_i\}$.

End Procedure

Figure 3.1: Procedure for computing the Z^*-ordering on rules.

The apparent circularity between κ^* and Z^* (Eqs. 3.20 and 3.21) is benign. Both functions can be computed recursively in an interleaved fashion, as shown in procedure Z^*_order (Fig. 3.1).

Theorem 3.13 *Given an MC set D, procedure Z^*_order computes the function Z^* defined by Eqs. 3.20 and 3.21.*

Corollary 3.14 *Given an MC set D, the function Z^* is unique.*

The function Z^* provides an economical way of storing the ranking κ^*, the space requirement is linear in the number of default rules. Still, in order to decide whether $\varphi \mathrel{\big|\!\sim} \psi$, we must check whether

$$\min_{\omega \models \varphi \wedge \psi} [\sum_{r_k \in D^-_\omega} Z^*(r_k)] < \min_{\omega \models \varphi \wedge \neg\psi} [\sum_{r_j \in D^-_\omega} Z^*(r_j)] \qquad (3.25)$$

47

holds, and the minimization required by Eq. 3.25 is NP-hard even for Horn expressions (see [9]).

3.4 Examples

Example 3.1 (Blue penguins) Consider the set $D_p = \{p \to \neg f, p \to b, b \to f\}$ of Example 2.2. An application of procedure Z^*_order will yield the ranking Z^*:

$$
\begin{aligned}
Z^*(b \to f) &= 1 \\
Z^*(p \to b) &= 2 \\
Z^*(p \to \neg f) &= 2
\end{aligned}
$$

Let bl be a new proposition denoting the color "blue". Note that since $\kappa^*(\varphi)$ depends solely on the Z^* of the rules violated in the preferred models of φ, and bl is a proposition that does not appear in any of the defaults in D_p, it follows that the defaults violated in the preferred models of $bl \wedge p \wedge \neg f$ ("blue penguins don't fly") are the same as those violated in the preferred possible worlds of $p \wedge \neg f$. We have:

$$
\begin{aligned}
\kappa^*(bl \wedge p \wedge \neg f) &= \min_{m \models bl \wedge p \wedge \neg f} \sum_{r_i \in D_{pM}^-} Z^*(r_i) = \kappa^*(p \wedge \neg f) = 1 \\
\kappa^*(bl \wedge p \wedge f) &= \min_{m \models bl \wedge p \wedge f} \sum_{r_i \in D_{pM}^-} Z^*(r_i) = \kappa^*(p \wedge f) = 2
\end{aligned}
$$

Thus, $\kappa^*(bl \wedge p \wedge \neg f) < \kappa^*(bl \wedge p \wedge f)$, which ratifies the conclusion $bl \wedge p \hspace{1mm}|\!\!\approx \neg f$ ("blue-penguins don't fly"). This conclusion follows directly from the rule of rational monotony (Eq. 3.6): Given that $p \to \neg f \in D_p$, $p \hspace{1mm}|\!\!\approx \neg f$ by the requirement of admissibility. Now, since any model $\omega \models p \wedge bl$ falsifies exactly the same rules as any model $\omega' \models p \wedge \neg bl$, $\kappa^*(p \wedge bl) = \kappa^*(p \wedge \neg bl)$ and $p \hspace{1mm}|\!\!\not\approx^* \neg bl$. Thus, by rational monotony $p \wedge bl \hspace{1mm}|\!\!\approx \neg f$. In general, we have the following proposition:

Proposition 3.15 *Let D be a set of defaults, and let p be a proposition not appearing in any of the defaults in D; then*

$$
p \wedge \phi \hspace{1mm}|\!\!\approx \sigma \quad \text{iff} \quad \phi \hspace{1mm}|\!\!\approx \sigma.
$$

Example 3.2 (Inheritance) In this example, we consider whether penguins, despite being an exceptional class of birds (with respect to flying), can inherit

48

other properties typical of birds. Consider the set $D_w = D_p \cup \{b \rightarrow w\}$.[12] The ranking Z^* remains identical to the one in Ex. 3.1 except for the new rule $Z^*(b \rightarrow w) = 1$. Note that since $\kappa^*(p \wedge w) = 1$ (in the most preferred model for $p \wedge w$, the default $b \rightarrow f$ is violated), while $\kappa^*(p \wedge \neg w) = 2$ (either both $b \rightarrow f$ and $b \rightarrow w$ or only $p \rightarrow b$ are violated in the most preferred models for $p \wedge \neg w$), we conclude $p \mathbin{\vert\!\sim} w$ (i.e., "penguins have wings").

It is instructive to compare the behavior of $\mathbin{\vert\!\sim}$ with the behavior of the rational closure of Lehmann [72] or, equivalently, system-Z [100]. System-Z selects the probability function P_ε^+ (among those admissible with D) that assigns to each world the highest possible probability (i.e., lowest rank) consistent with D.[13] Entailment is then defined in terms of the consequence relation induced by P_ε^*. A parallel can be drawn with the rational closure in terms of ranked-models. Rational closure also attempts to correct the conservative behavior of the *preferential entailment* [69] (which is essentially equivalent to probabilistic entailment) by restricting the set of consequence relations that define entailment. The resulting ranked-model and notion of entailment present the same properties that P_ε^+ and its consequence relation (this equivalence is formally shown in [48]). Given D_w (Ex. 3.2), $p \mathbin{\vert\!\sim} w$ will not follow from the rational closure of D_w: Once penguins are found to be exceptional birds with respect to flying, the consequence relation in the rational closure will regard penguins exceptional with respect to *any other* property of birds. The source of this counterintuitive behavior can be traced to the ranking function κ^+ (associated with P_ε^+) that sanctions this consequence relation defined by a pair of equations similar to Eqs. 3.20 and 3.21 [48, 100, 50]:

$$\kappa^+(\omega) = \max_{r_i \in D_\omega^-} Z^+(r_i) \tag{3.26}$$

where

$$Z^+(r_i) = \min_{M \models \varphi_i \wedge \psi_i} \kappa^+(\omega) \tag{3.27}$$

Whereas the maximum entropy approach uses "$\sum$" and tries to minimize the weighted sum of default violations, system-Z uses "max" and considers only the most significant default violated. Thus, a world in which a penguin has no wings ($\omega' \models p \wedge \neg w$) is no more surprising than one in which a penguin has wings; once the rule $b \rightarrow f$ is violated, the additional rule $b \rightarrow w$ violated in ω' does not alter κ^+.

The preference for worlds which violate less rules is a general property of the maximum entropy approach, and is made precise by the following proposition:

---

[12]To read "typically birds have wings".

[13]A summary of this strategy is presented in Chap. 4.

Proposition 3.16 *Let D be an MC set, D_M^- denote the set of defaults violated by model ω, and $D_{M'}^-$ denote the set of defaults violated by model ω'. If $D_M^- \subset D_{M'}^-$ then, $\kappa^*(\omega) < \kappa^*(\omega')$.*

In other words, if the set of defaults violated by a model ω is a proper subset of those violated by another model ω', then ω is strictly preferable to ω'. This *coherence* property seems natural, intuitive, and useful in applications such as fault diagnosis (see [110]). Suppose we know that "typically component p is not faulty" ($True \rightarrow \neg p$) and "typically component q is not faulty" ($True \rightarrow \neg q$). Given the observation that $p \vee q$ (either p or q are faulty), a reasonable conclusion to expect is that either p or q is faulty but not both. Indeed, since any model satisfying $[p \vee q] \wedge [p \wedge q]$ must violate a superset of the defaults violated by $[p \vee q] \wedge \neg[p \wedge q]$, we conclude $(p \vee q) \mathrel{\vphantom{|}\sim} (\neg p \vee \neg q)$, as expected.[14]

Example 3.3 (Independent properties) Consider $D_s = \{s \rightarrow w, s \rightarrow t\}$ standing for "typically Swedes are well-mannered" and "typically Swedes are tall". Since there is no explicit dependency between being well-mannered and being tall in D_s, a desirable default conclusion from D_s is $\neg t \wedge s \mathrel{\vphantom{|}\sim} w$ (i.e. "short-Swedes are well-mannered"). Again, this conclusion is sanctioned by $\mathrel{\vphantom{|}\sim}$ but not by the rational closure (or system-Z).

The $\mathrel{\vphantom{|}\sim}$ consequence relation is sensitive to the way in which the default rules are expressed. For example, had we encoded the information in D_s (Ex. 3.3) slightly differently, using $D_s' = \{s \rightarrow (w \wedge t)\}$ ("typically Swedes are well-mannered and tall"), we would no longer be able to conclude $s \wedge \neg t \mathrel{\vphantom{|}\sim} w$ ("typically Swedes who are not tall are well-mannered"). This sensitivity to the format in which rules are expressed seems at odds with one of the basic conventions of traditional logic, in which $a \rightarrow (b \wedge c)$ is regarded as shorthand for $a \rightarrow b$ and $a \rightarrow c$, and also stands in contrast with most other proposals for default reasoning (e.g. circumscription). However, this sensitivity might be useful for distinguishing fine nuances in natural discourse, treating w and t as two independent properties if expressed by two rules (i.e., "typically Swedes are tall" and "typically Swedes are well-mannered") and as related properties if expressed together (i.e., "typically Swedes are tall and well-mannered").

---

[14]This example is taken from [83], where the following question is posed: "Can the fact that we derive $\neg p \vee \neg q$ from $p \vee q$ when p, q are jointly circumscribed be explained in terms of probabilities close to 0 or 1?".

3.5 Non-Minimal-Core Sets

In the maximum entropy distribution, the constraint imposed by a default rule in D (see Eqs. 3.12 and 3.13) can be satisfied as either an equality (*active* constraint) or a strict inequality (*passive* constraint). The MC condition on D (Def. 3.11) guarantees that all these constraints are active. Once we relax this condition, not only does the process of finding a solution for the set of Eqs. 3.18 become more complex, but the resulting ranking may no longer represent a solution to the entropy maximization problem. This is because Eqs. 3.18 are the result of applying the Lagrange multipliers technique on the constraints represented by Eq. 3.12. This technique finds maxima on the boundaries defined by these constraints, blindly assuming that all constraints are satisfied as equalities (i.e., are active, see Appendix B), whereas all we are required to do is to satisfy the constraints of Eq. 3.12 with inequalities.

Another problem with using non MC sets D is that of redundant rules, that is, rules r that are already satisfied by the consequence relation $\mathrel{|\!\sim}$ induced by $D - \{r\}$ (see Prop. 3.12). It may be thought that these rules can be safely ignored and removed from the original knowledge base; however, in specifying a particular D, the user often intends for *all* rules in D to play an active role in *shaping* the consequence relation $\mathrel{|\!\sim}$. Overlooking this intention may lead to to counterintuitive results, as the following example demonstrates.[15]

Example 3.4 (Active set.) Consider the sets $D_a = \{a \to b, b \to c\}$ and $D_{a+} = D_a \cup \{a \to c\}$. Note that D_{a+} is not an MC set. If we run procedure Z^*_order on D_a, we find that the values $Z^*(a \to b)$ and $Z^*(b \to c)$ are both equal to 1, and that $a \mathrel{|\!\sim} c$ is in the consequence relation induced by D_a since $Z^*(a \to b)$ and $Z^*(b \to c)$ satisfy

$$\min_{\Omega^-_{a \to c}} [\sum_{r_k \in D^-_\omega} Z(r_k)] = 1 + \min_{\Omega^+_{a \to c}} [\sum_{r_j \in D^-_\omega} Z(r_j)]. \tag{3.28}$$

Thus, since the constraint imposed by the rule $a \to c$ in D_{a+} is satisfied by the maximum entropy solution to D_a, the two sets will have the same maximum entropy solution and the same consequence relation. Yet these two sets, D_a and D_{a+}, are not equivalent: While we do not expect $a \wedge \neg b \mathrel{|\!\sim} c$ to hold in D_a (the only possible "inference path" to c in D_a goes through b), we would like $a \wedge \neg b \mathrel{|\!\sim} c$ to be in any *reasonable* consequence relation induced by D_{a+} (where the rule $a \to c$ provides an alternative "path" to c).

---

[15]This example is a modified version of one originally suggested by Andrew Baker (personal communication).

The problem rests with using the equality $P_\varepsilon(c|a) = 1 - C \times \varepsilon$ as a constraint to the maximization process,[16] where in fact the constraint intended by the user is *stronger*, requiring a faster convergence of $P_\varepsilon^*(c|a)$ towards 1. The use of the Lagrange multipliers technique requires a commitment to a particular rate of convergence for each rule. Had we started with the insight that $\lim_{\varepsilon \to 0} P(c|a) = 1$ should be represented by $P(c|a) = 1 - C \times \varepsilon^2$ instead, the problem could have been solved using Eq. 3.18; all constraints in D_{a+} will be active, yielding

$$Z(a \to c) + \min_{\Omega_{a \to c}^-} [\sum_{\substack{r_k \in D_\omega^- \\ r_k \neq a \to c}} Z(r_k)] = 2 + \min_{\Omega_{a \to c}^+} [\sum_{r_j \in D_\omega^-} Z(r_j)], \qquad (3.29)$$

and the consequence relation $\,\mid\!\sim\,$ induced by D_{a+} will satisfy $a \wedge \neg b \,\mid\!\sim\, c$.

In general we could write Eq. 3.18 in terms of *slack variables* n_i for each constraint,

$$Z(r_i) + \min_{\Omega_{r_i}^-} [\sum_{\substack{r_k \in D_\omega^- \\ k \neq i}} Z(r_k)] = n_i + \min_{\Omega_{r_i}^+} [\sum_{r_j \in D_\omega^-} Z(r_j)] \qquad (3.30)$$

and seek the correct values of n_i that would render every rule active. Computationally, however, guessing the n_i's is not easier than guessing the set of active constraints.

We see that the advantages of MC sets are twofold: They guarantee convergence of procedure Z^*_order to a solution of Eq. 3.18, and this solution represents a solution to the entropy maximization problem. Unfortunately, to ensure these guarantees the expressiveness of the knowledge bases must be limited to MC, which may prevent us from specifying certain rule sets in the most natural way. It turns out that the class of knowledge bases where these guarantees hold is in fact wider than MC, since Eqs. 3.16 and 3.17 are valid as long as all rules are active. Consider the set D_p of Example 2.2,[17] augmented with the rule $p \to a$ ("typically penguins live in the Antarctic"). This set is not MC since any model falsifying $p \to a$ must falsify at least one other default in D_p. Nevertheless, all rules in $D_p \cup \{p \to a\}$ are active. Notice that $p \to a$ is *insensitive* to the Z-values associated with each of the rules in D_p, and vice versa. In other words, any change on the value of the Z associated with a rule in D_p will not affect the Z associated with $p \to a$.[18] Thus, we can compute the Z-values for D_p and $p \to a$

---

[16]Note that $P_\varepsilon(c|a) \geq 1 - C \times \varepsilon$ is equivalent to $C' \times \varepsilon \times P_\varepsilon(c \wedge a) \geq P_\varepsilon(\neg c \wedge a)$ for expressing $\lim_{\varepsilon \to 0} P(c|a) = 1$ as a constraint for the maximization process.

[17]Recall that $D_p = \{p \to \neg f, p \to b, b \to f\}$.

[18]The instance of Eq. 3.18 for $p \to a$ shows that the *min*-terms on both sides are identical and therefore can be canceled.

separately using two independent applications of procedure Z^*_order, one on D_p and the other on $p \to a$, and then combine the resulting Z's. We see that certain topological features of D permit its decomposition into a set of components belonging to MC, hence every rule will be active. The full characterization of databases in which all rules are active remains an open problem.

3.6 Discussion

Maximum entropy can be viewed as an extension of both p-entailment and the rational closure. Like p-entailment, it is based on infinitesimal probability analysis; and like rational closure, it is based on a unique ranking of possible worlds subject to constraints. In the rational closure, however, possible worlds are given the lowest rank that is consistently possible, and hence the rank of a model is equal to the rank of the most *crucial* rule violated by that model.[19] In contrast, maximum entropy ranks models according to the weighted sum of rule violations, and it is this difference that enables maximum entropy to sanction inheritance across exceptional subclasses, concluding that "penguins have wings" in Example 3.2 and that "short Swedes are well-mannered" in Example 3.3.

The ranking of rules in maximum entropy is reminiscent of abnormality preferences in prioritized circumscription [88], with the difference that the priorities assigned by maximum entropy are extracted automatically from the knowledge base and do not need the intervention of the user. This property is shared by Geffner's [36] *conditional entailment*, which also combines the virtues of the extensional and the conditional approaches. Conditional entailment maintains partial orders among rule priorities and among models, and it produces more acceptable inferences than maximum entropy in certain cases (see [36]), at the expense of a greater computational complexity and the loss of the underlying probabilistic semantics.

A weakness of the maximum entropy approach is that it stands at odds with the principle of causation.[20] If we first compute the maximum entropy distribution $P^*(X_1, \ldots, X_n)$ on a set of variables $X_1, \ldots, X_n$ and then consider one of their consequences Y, we may find that the maximum entropy distribution $P^*(X_1, \ldots, X_n, Y)$ constrained by the conditional probability $P(Y|X_1, \ldots, X_n)$ changes the probabilities of the X variables. For example, specifying the biases

---

[19]This ranking is called Z-rank in [100].

[20]This weakness, shared by many proposals for nonmonotonic reasoning [57], has required the introduction of special causal operators [116, 36]. Chapter 5 proposes a different approach to causality based on probabilistic considerations of independence.

of two coins yields a maximum entropy distribution in which the two coins are mutually independent. However, further specifying the probability with which an observer would respond to each of the four outcomes of these coins might yield a maximum entropy distribution in which the two coins are no longer independent of each other (see [97, pp. 463, 519], and [60]).[21] This stands contrary to our conception of causality in which the forecasting of future events does not alter beliefs about past events. This behavior prevents the maximum entropy approach from properly handling tasks such as the Yale shooting problem [57], where rules of causal character are given priority over other rules. Such priorities can be introduced in the κ-ranking system using a device called *stratification* (see Chap. 5), which forces κ to obey the so-called Markov condition: Knowing the present renders the future independent of the past. The role of maximum entropy in stratified rankings could then be to define preferred rankings under incomplete specification of causal influences: Out of all admissible rankings that conform to the stratification condition, choose those with the maximum entropy.

The maximum entropy formalism can be extended to admit defaults with variable strengths; each default in D can be annotated with a parameter δ_i, that indicates the firmness with which the default is believed.[22] Probabilistically, δ_i represents the slowest rate at which $P(\psi_i|\varphi_i)$ should be allowed to approach one as ε approaches zero. The constraints to the maximization process can be modified accordingly; thus, Eq. 3.12 will now read

$$C_i \times \varepsilon^{\delta_i} \times P_\varepsilon(\psi_i \wedge \varphi_i) \geq P_\varepsilon(\neg\psi_i \wedge \varphi_i), \tag{3.31}$$

and, given an MC set D, Eqs. 3.20 and 3.21 translate to

$$\kappa^*(\omega) = \sum_{r_i \in D_\omega^-} Z^*(r_i) \tag{3.32}$$

$$Z^*(r_i) = \min_{\omega \models \varphi_i \wedge \psi_i} \kappa^*(\omega) + 1 + \delta_i \tag{3.33}$$

where the Z^*-order for each rule can be computed using procedure Z^*_order.

This feature is very useful in domains such as circuit diagnosis where the analyst may feel strongly that failures are more likely to occur in one group of devices (e.g., multipiers) than in another (e.g., adders). For example, suppose that in addition to the information "typically component p is not faulty" and "typically component q is not faulty", we also know that component p is much

---

[21] Pearl attributes the discovery of this phenomenon to Norm Dalkey [97].

[22] An extension to system-Z [100] along these lines can be found in Chapter 4 (see also [50, 53]).

more likely to fail than component q. We can encode this information by specifying $D = \{True \xrightarrow{\delta_1} \neg p, True \xrightarrow{\delta_2} \neg q\}$ where $\delta_1 < \delta_2$. Thus, given that $p \vee q$ (either p or q are faulty) holds, we conclude that p is the faulty component with the failure $((p \vee q) \mathrel{\mid\!\sim} (p \wedge \neg q))$. If entailment is defined as the intersection of the $\mathrel{\mid\!\sim}$ consequence relations induced by all sets of values δ_i, then the resulting entailment relation will be supported by partial orders among rules and models, as in Geffner's [36] conditional entailment.

Chapter 4 studies such an extension to Pearl's system-Z [100], in which each $\varphi \rightarrow \psi$ is annotated with a positive integer δ denoting the degree of strength or firmness of the rule.

CHAPTER 4

Plausibility II: System-Z^+

4.1 Rankings as an Order-of-Magnitude Abstraction of Probabilities

Regardless of how we choose to interpret default statements, it is generally acknowledged that some defaults are stated with greater firmness than others. For example, the action-response defaults of the type "if Fred is shot with a loaded gun, Fred is dead" are issued with a greater conviction than persistence defaults of the type "if Fred is alive at time t, he is alive at $t + 1$". Moreover, the degree of conviction in this last statement should clearly depend on whether t is measured in years or seconds. In diagnosis applications, likewise, the analyst may feel strongly that failures are more likely to occur in one type of device (e.g., multipliers) than in another (e.g., adders). A language must be devised for expressing this valuable knowledge. Numerical probabilities or degrees of certainty have been suggested for this purpose, but if the full precision provided by numerical calculi is not necessary, an intermediate qualitative language might be more suitable.

This chapter proposes such a language in terms of the ranking functions introduced in Chapter 3 (Def. 3.7). This method of approximation gives rise to a semiqualitative calculus of uncertainty: Degrees of (dis)belief are ranked by nonnegative integers (corresponding perhaps to linguistic quantifiers such as "believable", "unlikely", "very rare", etc); retraction and restoration of beliefs conforms to Bayesian conditionalization.

One way of motivating ranking systems is to consider a probability distribution P defined over a set Ω of possible worlds and to imagine that an agent wishes to retain an order-of-magnitude approximate of P. The traditional engineering method of approximating P would be to express each numerical parameter (specifying P) in a base b representation, where b depends on the precision needed, and then omit all but the most significant figure from each expression.[1] All arithmetic

---

[1] Thus, given a number n and a basis b, its approximate would be the polynomial expression $a_0 * (b)^0 + a_1 * (b)^1 + a_2 * (b)^2 + \ldots$.

operations would then be performed on these approximate, single digit quantities, in lieu of the original parameters. The abstraction we advocate goes one step further. Instead of retaining the numerical value of the most significant figure, we retain only its *position*. The mechanics of this exercise is equivalent to, and can best be described by, a limit process where quantities are reqresented as polynomials in an infinitesimal number ε. These polynomials are added and multiplied precisely, but at the end we calculate the limit of the final results as ε goes to zero.

Imagine that the probability $P(\omega)$ is a polynomial function of some infinitesimal parameter ε, arbitrarily close to, yet larger than zero; for example, $P(\omega) = 1 - c_1 \varepsilon$ or $\varepsilon^2 - c_2 \varepsilon^4$.[2] Accordingly, the probabilities assigned to any subset of Ω represented by a logical formula φ, as well as all conditional probabilities $P(\psi|\varphi)$, will be rational functions of ε. We define the ranking function $\kappa(\psi|\varphi)$ as the power of the most significant ε-term in the expansion of $P(\psi|\varphi)$. In other words, $\kappa(\psi|\varphi) = n$ iff $P(\psi|\varphi)$ has the same order of magnitude as ε^n (see Def. 3.7).[3] Thus, instead of measuring probabilities on a scale from zero to one, we can imagine projecting probability measures onto a quantized logarithmic scale and then treating beliefs that map onto two different quanta as being of different orders of magnitude.

The following properties of ranking functions (left-hand side below) reflect, on a logarithmic scale, the usual properties of probability functions (right-hand side), with *min* replacing addition, and addition replacing multiplication:

$$\kappa(\varphi) = \min_{\omega \models \varphi} \kappa(\omega) \quad : \quad P(\varphi) = \sum_{\omega \models \varphi} P(\omega) \tag{4.1}$$

$$\kappa(\varphi) = 0 \text{ or } \kappa(\neg\varphi) = 0 \quad : \quad P(\varphi) + P(\neg\varphi) = 1 \tag{4.2}$$

$$\kappa(\psi \wedge \varphi) = \kappa(\psi|\varphi) + \kappa(\varphi) \quad : \quad P(\psi \wedge \varphi) = P(\psi|\varphi)P(\varphi) \tag{4.3}$$

Parameterizing a probability measure by ε and extracting the lowest exponent of ε as the measure of (dis)belief was proposed in [98] as a model of the process by which people abstract qualitative beliefs from numerical probabilities and accept them as tentative truths. For example, we can make the correspondence between linguistic quantifiers and ε^n depicted in Table 4.1 These approximations yield

---

[2] Probability functions parameterized on ε were called PPD's in Chapter 3. They are formally introduced in Definition 3.1, where the ε-functions were restricted to be analytical in $\varepsilon = 0$. The probability functions described above can also be viewed as the Taylor approximation of these PPD's.

[3] Spohn [120] was the first to study such ranking functions, which he named (non-probabilistic) ordinal condition function (OCF). His main motivation was to account for the dynamics of plain beliefs.

$P(\phi) = \varepsilon^0$	ϕ and $\neg\phi$ are believable	$\kappa(\phi) = 0$
$P(\phi) = \varepsilon^1$	$\neg\phi$ is believed	$\kappa(\phi) = 1$
$P(\phi) = \varepsilon^2$	$\neg\phi$ is strongly believed	$\kappa(\phi) = 2$
$P(\phi) = \varepsilon^3$	$\neg\phi$ is very strongly believed	$\kappa(\phi) = 3$
$\vdots$	$\vdots$	$\vdots$

Table 4.1: Linguistic quantifiers and ε^n.

a probabilistically sound calculus, employing integer addition, for manipulating the orders of magnitude of disbelief. The resulting formalism is governed by the following principles:

1. Each world is ranked by a non-negative integer κ representing the degree of surprise associated with finding such a world.

2. Each wff is given the rank of the world with the lowest κ (most normal world) that satisfies that formula.

3. Given a collection of facts ϕ, we say that σ follows from ϕ with strength δ if $\kappa(\sigma|\phi) > \delta$, or, equivalently, if the κ rank of $\phi \wedge \neg\sigma$ is at least $\delta + 1$ degrees above that of ϕ.

Principles 1 and 2 follow immediately from the semantics described above. Principle 3 says that σ is plausible given ϕ iff $P(\sigma|\phi) \geq 1 - c\varepsilon^{\delta+1}$, where P is the ε-parametrized probability associated with that particular ranking κ. This abstraction of probabilities matches the notion of *plain belief* in that it is deductively closed;[4] the drawback of this abstraction is that many small probabilities do not accumulate into a strong argument (as in the lottery paradox).

Reasoning using Principles 1 to 3 requires the specification of a complete ranking function. In other words, the knowledge base must be sufficiently rich to define the κ associated with every world ω.[5] Unfortunately, in practice, such specification is not readily available. We are usually given information in the form

---

[4]If A is believed and B is believed then, $A \wedge B$ is believed because $\kappa(\neg(A \wedge b)) > 0$ whenever $\kappa(\neg A) > 0$ and $\kappa(\neg B) > 0$. This deviates from the threshold conception of belief: if both $P(A)$ and $P(B)$ are above a certain threshold, $P(A \wedge B)$ may still be below that threshold.

[5]This is also the case with the OCF described in Spohn [120].

of statements such as "birds normally fly" which we interpret as $P(f|b) \geq 1 - \varepsilon$ or, equivalently, $\kappa(\neg f|b) > 0$, and no information whatsoever about the flying habits of red birds or non-birds. In this case, we still would like to conclude "red birds normally fly", even though the information given merely constraints κ to satisfy $\kappa(f \wedge b) < \kappa(\neg f \wedge b)$ ("it is less surprising to find a flying bird than a non-flying one"), and is not sufficient for defining a complete ranking function. Drawing plausible conclusions from such fragmentary pieces of information, requires additional inferential machinery to accomplish two functions: It should enrich the specification of the ranking function with the needed information, and it should operate directly on the specification sentences in the knowledge base, rather than on the rankings of worlds (which are too numerous to list). Such machinery is provided by a formalism called system-Z^+, described in this chapter, which accepts knowledge in the form of graded if-then rules and computes the plausibility of any given query by syntactic manipulation of these rules.

To accomplish these functions, system-Z^+ incorporates two principles in addition to those given above:

4. Each input rule "if φ then ψ (with strength δ)", written $\varphi \xrightarrow{\delta} \psi$, is interpreted as a constraint on the ranking κ, forcing every world in $\varphi \wedge \neg\psi$ to rank at least $\delta + 1$ degrees above the most normal world in φ, that is, $\kappa(\psi|\varphi) > \delta$.

5. Out of all rankings satisfying the constraints above, we adopt the ranking κ^+ that assigns each world the lowest possible (most normal) rank. Remarkably, this ranking will turn out to be unique.

Principle 4 is a straightforward generalization of the probabilistic reading of the rules, $P(\psi|\varphi) \geq 1 - c\varepsilon^{\delta+1}$. The parameter δ is an optional feature for the rule encoder that augments the expressiveness of the knowledge base by assigning strength to the rules. If δ is unspecified, it is assumed to be equal to zero, and rules are interpreted as $P(\psi|\varphi) \geq 1 - c\varepsilon$. A knowledge base with all $\delta = 0$ will be called a *flat* knowledge base. Remarkably, the addition of the δ's does not increase the computational complexity of query-answering and consistency testing.[6] Moreover, we shall see that even a flat knowledge base induces a natural priority on rules in order to respect specificity considerations (see Thm. 4.14).

Principle 5 reflects the assumption of maximal ignorance; unless compelled otherwise, assume every situation to be as normal as possible (or equivalently, no situation is more surprising than necessary). In contrast, the approach based on

---

[6] Both will require a polynomial number of propositional satisfiability tests.

maximum entropy (Chap. 3 selects the ranking κ^* that minimizes dependencies among propositions, to reflect only those implied by the rules in the knowledge base. As will be seen, the advantage of system-Z^+ is that the algorithm necessary for computing plausible conclusions is more efficient than the one for maximum entropy (Sec. 3.3). As in the case of maximum entropy, a key step in the procedure is the computation of a priority ordering Z^+ on the rules in the knowledge base.[7] Section 4.3 (after some preliminary definitions in Sec. 4.2), introduces a procedure that computes Z^+ in a polynomial number of propositional satisfiability tests and hence is tractable in applications permitting restricted languages, such as Horn expressions, network theories, or acyclic databases. Once the ordering Z^+ is known, the degree to which a given query is denied or confirmed can be computed in $O(log|\Delta|)$ satisfiability tests (where $|\Delta|$ is the number of rules in the knowledge base Δ). On the other hand, as shown in Section 4.7 and partially discussed in Sections 3.4 and 3.6, this computational advantage of system-Z^+ over maximum entropy, results in weakening the set of conclusions ratified by the system.

In Section 4.5, system-Z^+ is equipped with the capability to reason with *soft* evidence or imprecise observations. Such a capability is important when we wish to assess the plausibility of σ (using Prin. 3 above) but the context ϕ is not given with absolute certainty; all that can be ascertained is "ϕ is supported to a degree n". We propose two different strategies for computing a new ranking κ' from an initial one κ, given soft evidential report supporting a wff ϕ. The first strategy, named *J-conditionalization*, is based on Jeffrey's rule of conditioning [99]. It interprets the report as specifying that "all things considered", the new degree of disbelief for $\neg\phi$ should be $\kappa'(\neg\phi) = n$. The second strategy, named *L-conditionalization*, is based on the *virtual evidence* proposal described in [97, Chap. 2]. It interprets the report as specifying the desired *shift* in the degree of belief in ϕ, as warranted by that report alone and "nothing else considered". We show that both J and L-conditionalization have roughly the same complexity as ordinary conditionalization. Section 4.6 relates and compares system-Z^+ to the theory of belief revision in [3]. Finally, Section 4.7 summarizes the main results.

4.2 Preliminary Definitions: Rankings Revisited

Consider a set $\Delta = \{r_i \mid r_i = \varphi_i \xrightarrow{\delta_i} \psi_i, 1 \leq i \leq n\}$, where φ_i and ψ_i are propositional formulas, "$\rightarrow$" denotes a default connective as before, and δ_i is

---

[7]This priorities should be distinguished from the strengths δ's that are assigned to the rules by their author; priorities represent the interactions among the rules and reflect considerations such as specificity and relevance which are applicable to systems with flat knowledge bases.

a non-negative integer representing the degree of strength of rule r_i.[8] *Ranking functions* are defined as follows:

Definition 4.1 (Ranking) A *ranking function* κ is an assignment of non-negative integers to the elements in Ω, such that $\kappa(\omega) = 0$ for at least one $\omega \in \Omega$.
□

We extend this definition to induce rankings on wffs in accordance with the probabilistic interpretation of Eq. 4.1:

$$
\kappa(\varphi) = \begin{cases} \min_{\omega \models \varphi} \kappa(\omega) & \text{if } \varphi \text{ is satisfiable} \\ \infty & \text{otherwise} \end{cases} \tag{4.4}
$$

Similarly, following Eq. 4.3, we define the conditional ranking $\kappa(\psi|\varphi)$ for a pair of wffs φ and ψ as

$$
\kappa(\psi|\varphi) = \begin{cases} \kappa(\psi \wedge \varphi) - \kappa(\varphi) & \text{if } \kappa(\varphi) \neq \infty \\ \infty & \text{otherwise} \end{cases} \tag{4.5}
$$

Preferences are associated with lower κ, and *surprise* or *abnormality* with higher κ. Thus, $\kappa(\psi) < \kappa(\varphi)$ if ψ is preferred to φ in κ or, equivalently, if φ is more abnormal (surprising) than ψ. Intuitively, $\kappa(\psi|\varphi)$ stands for the degree of incremental *surprise* or *abnormality* associated with finding ψ to be true, given that we already know φ. The inequality $\kappa(\neg\psi|\varphi) > \delta$ means that given φ it would be surprising (i.e., abnormal) by at least $\delta+1$ additional ranks to find $\neg\psi$. Note that $\kappa(\neg\psi|\varphi) > \delta$ is equivalent to $\kappa(\varphi) + \delta < \kappa(\neg\psi \wedge \varphi)$ or $\kappa(\psi \wedge \varphi) + \delta < \kappa(\neg\psi \wedge \varphi)$, which is precisely the constraint on worlds we attribute to $\varphi \xrightarrow{\delta} \psi$.

Definition 4.2 (Consistency) A ranking κ is said to be *admissible* relative to a given Δ, iff

$$
\kappa(\varphi_i \wedge \psi_i) + \delta_i < \kappa(\varphi_i \wedge \neg\psi_i) \tag{4.6}
$$

(equivalently $\kappa(\neg\psi_i|\varphi_i) > \delta_i$) for every rule $\varphi_i \xrightarrow{\delta_i} \psi_i \in \Delta$. A knowledge base Δ is *consistent* iff there exists an admissible ranking κ relative to Δ.[9]
□

---

[8]For simplicity we skip the treatment of strict rules. The only necessary change required is in the conditions for admissibility in Def. 4.2. A strict rule $\phi \Rightarrow \sigma$ imposes the following admissibility constraint: $\kappa(\sigma \wedge \phi) < \kappa(\neg\sigma \wedge \phi)$ and $\kappa(\phi) < \infty$ (see Sec. 5.3, Eq. 5.8).

[9]Definition 4.2 represents the ranking equivalent of consistency and admissibility in Definition 3.2 (and Prp. 3.8) with a slight generatlization due to the new parameter δ. For reasons of simplicity I chose not to introduce a new term such as "κ-consistency". Also, as shown in Theorem 4.3 both notions are tested using the same procedure: Procedure Test_Consistency.
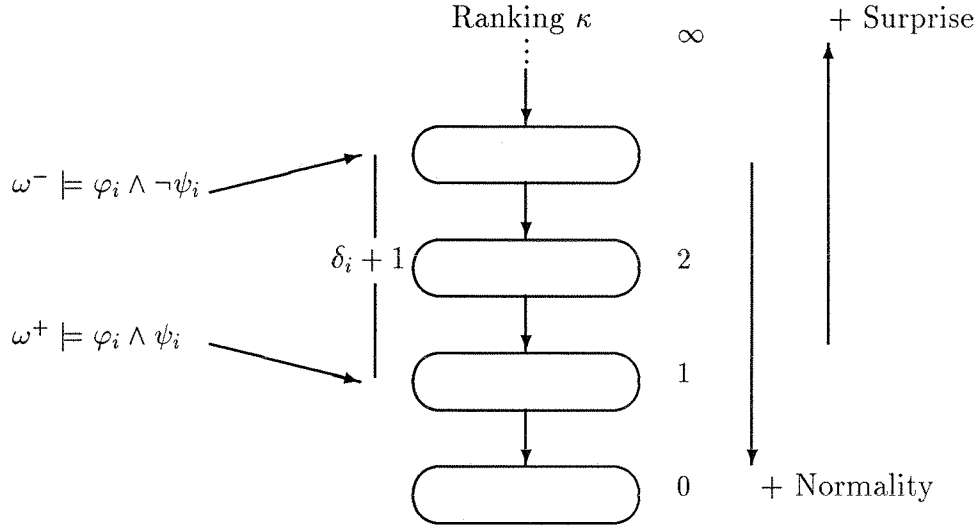
Figure 4.1: Consistency and rankings.

As expressed in Section 3.2 (when rankings were first introduced, see Def. 3.7 and Prop. 3.8), Eq. 4.6 echoes the usual interpretation of *default* rules [117], according to which ψ holds in all *minimal* models for φ. In our case, minimality is reflected in having the lowest rank, that is, the highest possible likelihood. Consistency guarantees that in every admissible ranking, each time we find a world ω^- violating a rule $\varphi_i \xrightarrow{\delta_i} \psi_i$, there must be a world ω^+ verifying $\varphi_i \xrightarrow{\delta_i} \psi_i$ such that $\kappa(\omega^+)$ must be at least $\delta_i + 1$ units less surprising than $\kappa(\omega^-)$ (see Fig. 4.1). In probabilistic terms, consistency guarantees that for every $\varepsilon > 0$, there exists a probability distribution P such that if $\varphi_i \xrightarrow{\delta_i} \psi_i \in \Delta$, then $P(\psi_i|\varphi_i) \geq 1 - c\varepsilon^{\delta_i+1}$.

Let $\bar{\Delta}$ denote a set of rules identical to Δ except that all the strengths δ_i are removed. Then,

Theorem 4.3 *A set Δ is consistent iff $\bar{\Delta}$ is p-consistent (Def. 2.2).*

Thus, consistency is independent of the δ_i strengths and we can use procedure Test_Consistency (Fig. 2.1) to test for consistency in a polynomial number of satisfiability tests. It is reassuring that once a knowledge base is consistent for one set of δ-assignments, it will be consistent with respect to any such assignment, which means that the rule author has the freedom to modify the δ's without fear of forming an inconsistent knowledge base.

4.3 Plausible Conclusions: The Z^+-Rank

Given a set Δ, each admissible ranking κ induces a consequence relation $\mathrel{\vdash\!\!\sim}_{K}$, where $\phi \mathrel{\vdash\!\!\sim}_{K} \sigma$ iff $\kappa(\sigma \wedge \phi) < \kappa(\neg\sigma \wedge \phi)$. A straightforward way to declare σ as a plausible conclusion of Δ given ϕ would be to require $\phi \mathrel{\vdash\!\!\sim}_{K} \sigma$ in all κ admissible with Δ. However, this will result in an entailment relation equivalent to p-entailment which was shown to be too conservative (Chap. 3). Thus, similar to the approach taken in Chapter 3, we select a distinguished admissible ranking, κ^+, and declare σ as a plausible conclusion of Δ given ϕ, written $\phi \mathrel{\vdash\!\!\sim}_{+} \gamma$, iff $\kappa^+(\phi \wedge \sigma) < \kappa^+(\phi \wedge \neg\sigma)$.[10]

The ranking κ^+ assigns to each world the lowest possible rank permitted by the admissibility constraints of Eq. 4.6. We will first introduce a syntactic definition of κ^+ and then show that it satisfies the desired minimality condition.

Definition 4.4 (The ranking κ^+) Let $\Delta = \{r_i \mid r_i = \varphi_i \overset{\delta_i}{\to} \psi_i\}$ be a consistent set of rules. κ^+ is defined as follows:

$$\kappa^+(\omega) = \begin{cases} 0 & \text{if } \omega \text{ does not falsify any rule in } \Delta \\ \max_{\omega \models \varphi_i \wedge \neg\psi_i}[Z^+(r_i)] + 1 & \text{otherwise} \end{cases} \qquad (4.7)$$

where $Z^+(r_i)$ is a *priority* ordering on rules, defined by

$$Z^+(r_i) = \min_{\omega \models \varphi_i \wedge \psi_i}[\kappa^+(\omega)] + \delta_i. \qquad (4.8)$$

$\square$

Eqs. 4.7 and 4.8 can be viewed as two coupled equations; one defines κ^+ in terms of Z^+, the second defines Z^+ in terms of κ^+. Figure 4.2 presents an effective procedure, procedure Z^+_order, for computing Z^+ from Δ. The significance of Eq. 4.8 is that the priorities function Z^+ constitute an economical way of encoding the ranking κ^+, linear in the size of Δ, from which the κ^+ of any world ω can be computed by searching the highest Z^+ rule violated by ω in a logarithmic number (on the number of rules in Δ) of satisfiability tests. The resulting consequence relation $\mathrel{\vdash\!\!\sim}_{+}$ and its associated reasoning procedures are called system-Z^+.

We next show (Thm. 4.7 and Cor. 4.8) that Eqs. 4.7 and 4.8 define a unique admissible ranking function κ^+ that is minimal in the following sense: Any other admissible ranking function must assign a higher ranking to at least one world and

---

[10]If we are concerned with the strength δ with which the conclusion is endorsed, then $\phi \mathrel{\vdash\!\!\sim}_{+}^{\delta} \gamma$ iff δ is the lowest (positive) integer I satisfying $\kappa^+(\phi \wedge \sigma) + I < \kappa^+(\phi \wedge \neg\sigma)$.

a lower ranking to none. To make the result formal, we introduce the following definitions:

Definition 4.5 (Minimal ranking) A ranking function κ is said to be *minimal* if every other admissible ranking κ' satisfies $\kappa'(\omega) > \kappa(\omega)$ for at least one possible world ω.
□

Definition 4.6 (Compact rankings) An admissible ranking κ is said to be *compact* if, for every ω' any ranking κ' satisfying

$$\begin{aligned}
\kappa'(\omega) &= \kappa(\omega) & \omega \neq \omega' \\
\kappa'(\omega) &< \kappa(\omega) & \omega = \omega'
\end{aligned}$$

is *not* admissible.
□

Theorem 4.7 *Every consistent Δ has a unique compact ranking given by κ^+.*

Corollary 4.8 *Every consistent Δ has a unique minimal ranking given by κ^+.*

Note the similarity between k^+ (Eq. 4.7) and the ranking k^* associated with the maximum entropy approach (reproduced below)

$$\kappa^*(\omega) = \begin{cases} 0 & \text{if } \omega \text{ does not falsify any rule in } \Delta, \\[2mm] \sum_{\omega \models \varphi_i \wedge \neg \psi_i} [Z^*(r_i)] + 1 & \text{otherwise.} \end{cases} \tag{4.9}$$

While $\kappa^+(\omega)$ is defined by the maximum-priority rule violated in ω, $\kappa^*(\omega)$ depends on the summation of these priorities. This difference will have implications for both the computational complexity and the quality of conclusions that these two proposals sanction.

The computation of the Z^* priorities and the query-answering procedures for the maximum entropy approach has been proven to be NP-hard even for Horn clauses (see [9]). In contrast, the computation of Z^+ using Procedure Z^+_order can be accomplished $O(|\Delta|^2 \times \log |\Delta|)$ satisfiability tests (Thm 4.12). The procedure for computing Z^+ is presented in Figure 4.2, and is very similar to the one in Figure 3.1. Some of the steps in procedure Z^+_order invoke a test of *toleration* (Def. 2.3). A rule $\phi \xrightarrow{\delta} \sigma$ is *tolerated* by Δ if the wff $\phi \wedge \sigma \bigwedge_i \varphi_i \supset \psi_i$ is satisfiable (where i ranges over all rules in Δ).

Theorem 4.9 establishes the correctness of procedure Z^+_order, while Lemmas 4.10 and 4.11 and Theorem 4.12 determine its (polynomial) complexity.

Procedure Z^+_order
Input: A consistent knowledge base Δ. Output: Z^+-ranking on rules.

1. Let Δ_0 be the set of rules tolerated by Δ, and let $\mathcal{R}Z^+$ be an empty set.

2. For each rule $r_i = \varphi_i \xrightarrow{\delta_i} \psi_i \in \Delta_0$, set $Z(r_i) = \delta_i$ and $\mathcal{R}Z^+ = \mathcal{R}Z^+ \cup \{r_i\}$.

3. While $\mathcal{R}Z^+ \neq \Delta$, do:

   (a) Let Δ^+ be the set of rules in $\Delta' = \Delta - \mathcal{R}Z^+$ tolerated by Δ'.

   (b) For each $r : \phi \xrightarrow{\delta} \sigma \in \Delta^+$, let Ω_r denote the set of models for $\phi \wedge \sigma$ that do not violate any rule in Δ'; compute

   $$Z(r) = \min_{\omega_r \in \Omega_r} [\kappa(\omega_r)] + \delta \qquad (4.10)$$

   where $\kappa(\omega_r) =$

   $$\max_{r_j \in \mathcal{R}Z^+} \{ Z(r_j) \mid \omega_r \models \varphi_j \wedge \neg \psi_j \} + 1 \qquad (4.11)$$

   and $r_j : \varphi_j \xrightarrow{\delta_j} \psi_j \in \mathcal{R}Z^+$.

   (c) Let r^* be a rule in Δ^+ having the lowest Z; set $\mathcal{R}Z^+ = \mathcal{R}Z^+ \cup \{r^*\}$.

End Procedure

Figure 4.2: Procedure for computing the Z^+-ordering on rules.

Theorem 4.9 *The function Z computed by procedure Z^+_order satisfies Def. 4.4, that is $Z = Z^+$.*[11]

Lemma 4.10 *Let $\Delta = \{r_i \mid r_i = \varphi_i \xrightarrow{\delta_i} \psi_i\}$ be a consistent knowledge base in which rules are sorted according a priority function $Z(r_i)$. Let $\kappa(\omega)$ be defined as in Eq. 4.7:*

$$\kappa(\omega) = \begin{cases} 0 & \text{if } \omega \text{ does not falsify any rule in } \Delta, \\ \\ \max_{\omega \models \varphi_i \wedge \neg\psi_i}[Z(r_i)] + 1 & \text{otherwise.} \end{cases} \quad (4.12)$$

Then, for any wff ϕ, $\kappa(\phi)$ can be computed in $O(\log|\Delta|)$ propositional satisfiability tests.

The idea is to perform a binary search on Δ to find the lowest $Z(r)$ such that there is a model for ϕ that does not violate any rule r' with priority $Z(r') \geq Z(r)$. This is done by dividing Δ into two roughly equal sections: top–half (r_{mid} to r_{high}) and bottom–half (r_{low} to r_{mid}). A satisfiability test on the wff $\alpha = \phi \wedge \bigwedge_{j=mid}^{j=n} \varphi_j \supset \psi_j$ decides whether the search should continue (in a recursive fashion) on the bottom-half or top-half.

Lemma 4.11 *The value of $Z(\phi \xrightarrow{\delta} \sigma)$ in Eq. 4.10 can be computed in $O(\log|\mathcal{R}Z^+|)$ satisfiability tests.*

Let Δ' in Step 3(a) of procedure Z^+_order be equal to $\{\varphi_i \xrightarrow{\delta_i} \psi_i\}$; the computation of Eq. 4.10 is equivalent to computing the κ of the wff $\sigma \wedge \phi \bigwedge_i \varphi_i \supset \psi_i$ where i ranges over all the rules in Δ' by performing the binary search on the set $\mathcal{R}Z^+$.

Theorem 4.12 *Given a consistent Δ, the computation of the ranking Z^+ requires $O(|\Delta|^2 \times \log|\Delta|)$ satisfiability tests.*

Computing Eq. 4.10 in Step 3(b) can be done in $O(\log|\mathcal{R}Z^+|)$ satisfiability tests according to Lemma 4.11,[12] and, since it will be executed at most $O(|\Delta|)$ times, it requires a total of $O(|\Delta| \times \log|\Delta|)$. Loop 3 is performed at most $|\Delta| - |\Delta_0|$

---

[11]Note that Eqs. 4.10 and 4.11 correspond to Eqs. 4.8 and 4.7 in Def. 4.4.

[12]Note that we need $\mathcal{R}Z^+$ to be sorted, nondecreasingly, with respect to the priorities Z. This requires that the initial values inserted in $\mathcal{R}Z^+$ in Step 2 of procedure Z^+_order be sorted taking $O(|\Delta_0|^2)$ data comparisons and that the new Z-value in Step 3(c) be inserted in the right place taking $O(|\mathcal{R}Z^+|)$ data comparisons. We are assuming that the cost of each of these operations is much less than that of a satisfiability test.

times, hence the whole computation of the priorities Z^+ on rules requires a total of $O(|\Delta|^2 \times \log |\Delta|)$ satisfiability tests.[13]

Once Z^+ is known, determining the strength δ with which an arbitrary query σ is confirmed, given the information ϕ, requires $O(\log |\Delta|)$ satisfiability tests: First $\kappa^+(\phi \wedge \sigma)$ and $\kappa^+(\phi \wedge \neg\sigma)$ are computed, using a binary search as in Lemma 4.10. Then, these two values are compared and the difference is equated with the strength δ. Clearly, if the rules in Δ are of Horn form, computing the priority ranking Z^+ and deciding the plausibility of queries $(\phi \vmodels{\delta} \sigma)$ can be done in polynomial time [25].

In the special case of a *flat* Δ, that is, all δ's $= 0$, the procedure reduces to the following steps: First, identify all rules $r_i : \varphi_i \rightarrow \psi_i$ in Δ for which the formula

$$\varphi_i \wedge \psi_i \bigwedge_{j \neq i, r_j \in \Delta} \varphi_j \supset \psi_j \tag{4.13}$$

is satisfiable. Next, assign to these defaults priority $Z^+ = 0$, remove them from Δ, and repeat the process, assigning to the next set of defaults the priority $Z^+ = 1$, then $Z^+ = 2$, and so on. Once Z^+ is known, the rank κ^+ of any wff ϕ is given by $\kappa^+(\phi) = $ minimum i, such that

$$\phi \bigwedge_{j : Z^+(r_j) \geq i} \varphi_j \supset \psi_j \tag{4.14}$$

is satisfiable. This special case of a flat Δ constitutes system-Z as introduced by Pearl [100].

An important result implied by Eqs. 4.13 and 4.14 gives a method of constructing a propositional theory $Th(\phi)$ that implies all the conclusions σ that plausibly follow from a given evidence ϕ, that is, $\phi \vmodels{} \sigma$. Such a theory is given by the formula

$$Th(\phi) = \bigwedge_{i : Z^+(r_i) \geq \kappa^+(\phi)} \varphi_i \supset \psi_i. \tag{4.15}$$

This is somewhat reminiscence of Brewka's [16] and Poole's [104] idea of constructing preferred subtheories that are maximally consistent with the context ϕ. Here , the construction is more cautious; it stops as soon as all rules of priority $Z^+ \geq \kappa^+(\phi)$ are included in the theory. Ways of completing the construction were proposed by Boutilier [15] (see discussion in Sec. 4.7). Note, however, that in contrast to Brewka's and Poole's proposals, our priorities are computed automatically from the knowledge base.

---

[13]The complexity of the rest of the steps in the procedure is bounded by $O(|\Delta|)$ satisfiability tests.

4.4 Examples

The following examples illustrate properties of the κ^+-ranking and the use of δ to impose priorities among defaults. Example 4.1 shows how the specificity-based preferences in Example 2.2 are established and maintained by the κ^+-ranking, freeing the rule encoder from such considerations.[14] In Example 4.2, the strengths δ are used to establish preferences when specificity relations are not available.

Example 4.1 (Irrelevance and specificity) Consider the following set of rules taken from Example 2.2:

$$r_1: \ b \xrightarrow{\delta_1} f$$

$$r_2: \ p \xrightarrow{\delta_2} b$$

$$r_3: \ p \xrightarrow{\delta_3} \neg f$$

$$r_4: \ f \xrightarrow{\delta_4} a$$

standing for r_1:"birds fly", r_2:"penguins are birds", r_3:"penguins don't fly", and r_4:"flying things are airborne". The Z^+-ordering is computed as follows: Since both r_1 and r_4 are tolerated by all the rules in the knowledge base, $Z^+(r_1) = \delta_1$ and $Z^+(r_4) = \delta_4$. Any κ^+-minimal world verifying r_2 or r_3 must violate r_1; therefore, following procedure Z^+_order, $Z^+(r_2) = \delta_1 + \delta_2 + 1$ and $Z^+(r_3) = \delta_1 + \delta_3 + 1$. The first column of Table 4.2 contains some queries, the second contains p-entailed conclusions, and the last contains conclusions entailed by system-Z^+. The reason system-Z+ concludes that "red birds fly" ($r \wedge b \mathrel{\mid\!\sim}_+ f$ is

Queries	p-entailment	system-Z^+
$(p \wedge b, f)$ – "Do penguin-birds fly?"	NO	NO
$(r \wedge b, f)$ – "Do red birds fly?"	Possibly	YES
(b, a) – "Are birds airborne?"	Possibly	YES

Table 4.2: Plausible conclusions in Example 3.1.

as follows: Since r is a proposition that does not appear in the knowledge base,

14A general formalization of this behavior is Theorem 4.14.

68

any rule violated by a world $\omega \models b \wedge f$ is also violated by a world $\omega' \models b \wedge r \wedge f$. Thus, conclusions in system-Z^+ are unperturbed by irrelevant propositions. In general, we have:[15]

Proposition 4.13 *Let Δ be a consistent set of defaults, and let p be a proposition not appearing in any of the defaults in Δ; then $p \wedge \phi \mathrel{\vDash_{\overline{+}}} \sigma$ iff $\phi \mathrel{\vDash_{\overline{+}}} \sigma$.*

Another interesting conclusion sanctioned by $\mathrel{\vDash_{\overline{+}}}$ that is not p-entailed is "birds are normally airborne". Note that this inference reflects a limited form of rule chaining (not present in p-entailment).

Finally, as in p-entailment, the priorities Z^+ recognize that r_3 is *more specific* than r_1 and sanctions "a penguin-bird does not fly". Note that the preference for r_3 over r_1 is established independently of the initial δ's assigned to these rules. In the knowledge base above, the priority of r_3 ("typically penguins do not fly") was adjusted from δ_3 to $\delta_1 + \delta_3 + 1$, so as to supersede δ_1, the priority of the conflicting rule "typically birds fly". As a result of such adjustments, the relative importance of rules is maintained throughout the system, and compliance with specificity-type constraints is automatically preserved. This is made precise in the following theorem.

Theorem 4.14 *Let $r_1 : \varphi \xrightarrow{\delta_1} \psi$ and $r_2 : \phi \xrightarrow{\delta_2} \sigma$ be two rules in a consistent Δ such that*

1. $\varphi \mathrel{\vDash_{p}} \phi$ (i.e., φ is more specific than ϕ).

2. There is no model satisfying $\varphi \wedge \psi \wedge \phi \wedge \sigma$ (i.e., r_1 conflicts with r_2).

Then $Z^+(r_1) > Z^+(r_2)$ independent of the values of δ_1 and δ_2.

In other words, the Z^+-ordering guarantees that features of more specific contexts override conflicting features of less specific contexts.

Example 4.2 (Belief strength) Consider the following knowledge base (a subset of Example 2.3):

$r_1: q \xrightarrow{\delta_1} p$

$r_2: r \xrightarrow{\delta_2} \neg p$

---

[15]Note that this proposition is system-Z^+ counterpart to maximum entropy Prp. 3.15.

standing for r_1:"typically Quakers are pacifists" and r_2:"typically Republicans are not pacifists". Since each rule is tolerated by the other, the Z^+ of each rule is equal to its associated δ: $Z^+(r_1) = \delta_1$ and $Z^+(r_2) = \delta_2$. Given an individual, say Nixon, who is both a Republican and a Quaker, the decision of whether Nixon is a pacifist will depend on whether δ_1 is larger than, less than, or equal to δ_2. This is because any model ω_{rqp} for Quakers, Republicans, and pacifists must violate r_2, and consequently $\kappa^+(\omega_{rqp}) = \delta_2$, while any model $\omega_{rq\neg p}$ for Quakers, Republicans, and non-pacifists must violate r_1, that is, $\kappa^+(\omega_{rq\neg p}) = \delta_1$. In this case the decision to prefer one world over the other does not depend on specificity considerations but rather on whether the rule encoder believes that religious convictions carry more weight than political affiliations.

The main shortcomings of system-Z^+ are discussed in Sections 3.4 and 4.7.

4.5 Belief Change, Soft Evidence, and Imprecise Observations

So far, a query $\phi \vdash^{\delta}_{\tilde{+}} \sigma$ was defined as a pair of Boolean formulas (ϕ, σ), where ϕ (the *context*) stands for the set of observations at hand and σ (the *target*) stands for the conclusion whose belief we wish to confirm, deny, or assess. A query (ϕ, σ) would be answered in the affirmative if σ was found to hold in all minimally ranked models of ϕ, and the degree of belief in σ would be given by $\kappa(\neg\sigma \wedge \phi) - \kappa(\sigma \wedge \phi)$.

In many cases, however, the queries we wish to answer cannot be cast in this format, because our set of observations is not precise enough to be articulated as a crisp Boolean formula. For example, assume that we are throwing a formal party and our friends Mary and Bill are invited. However, judging from their previous behavior, we believe "if Mary goes to the party, Bill will stay home (with strength δ)", written $M \xrightarrow{\delta} \neg B$. Now assume that we have a strong hunch (with degree K) that Mary will go to the party (perhaps because she is extremely well dressed and is not consulting the movie section in the *Times*) and we wish to inquire whether Bill will stay home. It would be inappropriate to query the system with the pair $(M, \neg B)$, because the context M has not been established beyond doubt. The difference could be critical if we have arguments against "Bill staying home", (e.g., he was seen renting a tuxedo). A flexible system should allow the user to assign a degree of belief to each observational proposition in the context ϕ and proceed with analyzing their rational consequences. Thus, a query should consist of a tuple like $(\phi_1, K_1; \phi_2, K_2; \ldots, \phi_m, K_m : \sigma)$, where each K_i measures

the degree to which the contextual proposition ϕ_i is supported by evidence.[16]

At first glance it might seem that system-Z^+ would automatically provide such a facility through the use of variable-strength rules. For example, to express the fact that Mary is believed to be going to the party, we can perhaps use a *dummy* rule $Obs_1 \xrightarrow{K} M$ (stating that if Mary meets the set of observations Obs_1, then Mary is believed to be going to the party) and then add the proposition Obs_1 to the context part of the query, to indicate that Obs_1 has taken place.

This proposal has several shortcomings, however. First, in many systems it is convenient to treat if–then rules as a stable part of our knowledge, unperturbed by observations made about a particular individual or in any specific set of circumstances. This permits us to compile rules into a structure that allows efficient processing over a long stream of queries. Adding query-induced rules to the knowledge base will neutralize this facility.

Second, rules and observations combine differently: The latter should accumulate, the former do not. For example, if we have two rules $a \xrightarrow{\delta_1} c$ and $b \xrightarrow{\delta_2} c$ and we observe a and b, system-Z^+ would believe c to a degree $\max(\delta_1, \delta_2)$. However, if a and b provide two independent reasons for believing c, the two observations together should endow c with a belief that is stronger than any one component in isolation. To incorporate such cumulative pooling of evidence, we must encode the assumption that a and b are conditionally independent (given c), which is not automatically embodied in system-Z^+.[17]

To avoid these complications, the method we propose treats imprecise observations by invoking specialized conditioning operators, unconstrained by a rule's semantics. We distinguish between two types of evidential reports:

1. Type-J: "All things considered," our current belief in ϕ should *become J*.

2. Type-L: "Nothing else considered," our current belief in ϕ should *shift by L*.

---

[16]We remark that evidence in this dissertation is regarded as setting the context of a query and not as a modifier of the knowledge in Δ. Statistical methods for accomplishing the latter task are explored by Bacchus [6].

[17]The assumptions of conditional independence among converging rules is embodied in the formalism of maximum entropy (see Chapter 3 and [47]), as well as in the causal interpretation of Chapter 5.

4.5.1 Type-J: All Things Considered

Let ϕ be the wff representing the event whose belief we wish to update so that $\kappa'(\neg\phi) = J$ (and, consequently, $\kappa'(\phi) = 0$).[18] In order to compute $\kappa'(\psi)$ for every wff ψ, we rely upon Jeffrey's rule of conditioning [99]. Jeffrey's rule is based on the assumption that when an agent reports that an observation changed her degree of belief in ϕ, such observation does not normally change the *conditional degree of belief* in any propositions conditional on the evidence ϕ or on the evidence $\neg\phi$ [99]. Thus, letting P and P' denote the agent's probability distribution before and after the observation respectively, we have[19]

$$P'(\psi|\phi) = P(\psi|\phi) \text{ and } P'(\psi|\neg\phi) = P(\psi|\neg\phi), \tag{4.16}$$

which leads to Jeffrey's rule,

$$P'(\psi) = P(\psi|\phi)P'(\phi) + P(\psi|\neg\phi)P'(\neg\phi). \tag{4.17}$$

Translated into the language of rankings (using Eqs. 4.1–4.3), Eq. 4.17 yields

$$\kappa'(\psi) = \min[\kappa(\psi|\phi) + \kappa'(\phi); \kappa(\psi|\neg\phi) + \kappa'(\neg\phi)], \tag{4.18}$$

which offers a convenient way of computing $\kappa'(\psi)$ once we specify $\kappa'(\phi) = 0$ and $\kappa'(\neg\phi) = J$. Eq. 4.18 assumes an especially attractive form when computing the κ' of a world ω:

$$\kappa'(\omega) = \begin{cases} \kappa(\omega|\phi) + \kappa'(\phi) & \text{if } \omega \models \phi \\ \kappa(\omega|\neg\phi) + \kappa'(\neg\phi) & \text{if } \omega \models \neg\phi \end{cases} \tag{4.19}$$

Eq. 4.19 corresponds exactly to the α-conditionalization proposed in Spohn [120] (Def. 6, p. 117), with $\alpha = J$. If $\kappa'(\neg\phi) = \infty$, this process is equivalent to ordinary Bayesian conditionalization, since $k'(\omega) = k(\omega|\phi)$ if $\omega \models \phi$ and $\kappa'(\omega) = \infty$ otherwise. Note, however, that in general this conditionalization is not commutative; if ϕ_1 and ϕ_2 are mutually dependent (i.e., $\kappa(\phi_2|\phi_1) \neq \kappa(\phi_2)$),[20] the order in which we establish $\kappa(\neg\phi_1) = J_1$ and $\kappa(\neg\phi_2) = J_2$ might make a difference in our final

---

[18]This is an immediate consequence of the semantics for rankings and corresponds to the normalization in probability theory (see Eq. 4.2).

[19]Eq. 4.16 is known as the *J–condition* [99].

[20]This condition mirrors probabilistic dependence, namely, $P(\phi_2|\phi_1) \neq P(\phi_2)$.

belief state.[21] This is not surprising since in the "all things considered interpretation" the last report is presumed to summarize *all* previous observations.

4.5.2 Type-L Reports: Nothing Else Considered

L-conditionalization is appropriate for evidential reports of the type "a new evidence was obtained which, by its own merit, would *support* ϕ to degree L." Unlike J-conditionalization, the degree L now specifies *changes* in the belief of ϕ, not the absolute value of the final belief in ϕ. As in the case of type-J reports, we assume that in naming ϕ as the direct beneficiary of the evidence, the intent is to convey the assumption of conditional independence, as formulated in Eq. 4.17. Next, following the method of *virtual conditionalization* [97], we assume that the degree of support L characterizes the likelihood-ratio $\lambda(\phi)$ associated with some undisclosed observation *Obs*:

$$\lambda(\phi) = \frac{P(Obs|\phi)}{P(Obs|\neg\phi)}, \tag{4.20}$$

which governs the updates via the product rule

$$\frac{P'(\phi)}{P'(\neg\phi)} = \frac{\lambda(\phi)P(\phi)}{P(\neg\phi)}. \tag{4.21}$$

Translated into the language of rankings, this assumption yields

$$\kappa'(\phi) - \kappa'(\neg\phi) = \kappa(\phi) - \kappa(\neg\phi) - L \tag{4.22}$$

and, since either $\kappa'(\phi)$ or $\kappa'(\neg\phi)$ must be zero, we obtain

$$\kappa'(\phi) = \max[0; \kappa(\phi) - \kappa(\neg\phi) - L], \tag{4.23}$$

$$\kappa'(\neg\phi) = \max[0; \kappa(\neg\phi) - \kappa(\phi) + L]. \tag{4.24}$$

We see that the effect of L-conditionalization is to shift the difference between the degrees of disbelief in ϕ and $\neg\phi$ by the specified amount L. Once $\kappa'(\phi)$ is known, Jeffrey's rule (Eq. 4.18) can be used to compute the $\kappa'(\sigma)$ for an arbitrary

---

[21]Spohn ([120], p. 118) has acknowledged the desirability of commutativity in evidence pooling but has not stressed that α-conditionalization commutes only in a very narrow set of circumstances (partially specified by his Thm. 4). These circumstances require that successive pieces of evidence support only propositions that are relatively independent — the truth of one proposition should not imply a belief in another. Shenoy [114] has corrected this deficiency by devising a commutative combination rule that behaves like to L-conditioning.

wff σ yielding

$$\kappa'(\sigma) = \begin{cases} \min[\kappa(\psi|\phi) + \kappa(\phi) - \kappa(\neg\phi) - L; \kappa(\psi|\neg\phi)] \\ \min[\kappa(\psi|\phi); \kappa(\psi|\neg\phi) + \kappa(\neg\phi) + L - \kappa(\phi)] \\ \min[\kappa(\psi|\phi); \kappa(\psi|\neg\phi)] \end{cases} \qquad (4.25)$$

depending on whether $\kappa(\neg\phi) + \kappa(\phi)$ is less than, greater than, or equal to L. This expression takes the following form for $\kappa'(\omega)$:

$$\kappa'(\omega) = \begin{cases} \kappa(\omega|\phi) + \max[0; \kappa(\phi) - \kappa(\neg\phi) - L] & \text{if } \omega \models \phi, \\ \kappa(\omega|\neg\phi) + \max[0; \kappa(\neg\phi) - \kappa(\phi) + L] & \text{if } \omega \models \neg\phi. \end{cases} \qquad (4.26)$$

As in J-conditionalization, if $L = \infty$ then $\kappa'(\omega) = \kappa(\omega|\phi)$. For the general case, we can see that the effect of L-conditionalization is to shift *downward* the κ of all worlds that are models of the supported proposition ϕ relative to the κ of all worlds that are not models for ϕ. However, unlike J-conditionalization, the net relative shift is constant and is equal to L, independent of the initial value of $\kappa(\phi)$. It is easy to verify that L-conditionalization is commutative (as is its probabilistic counterpart, see Eq. 4.21), and hence it permits a recursive implementation in the case of multiple evidence.

We can illustrate these updating schemes through the party example consisting of the single rule $r_m : M \overset{4}{\to} \neg B$ ("if Mary goes to the party, then Bill will not go"). A trivial application of procedure Z^+_order yields $Z^+(r_m) = 4$, and using Eqs. 4.4 and 4.7 we find $\kappa(x) = 0$ for every proposition x, except $x = B \wedge M$, for which $\kappa^+(M \wedge B) = 5$. This means that we have no reason to believe that either Mary or Bill will go to the party, but we are pretty sure that both of them will not show up. Now suppose we see that Mary is very well dressed, and this observation makes our belief in M increase to 3, that is, $\kappa^{+'}(\neg M) = 3$. As a consequence, our belief in Bill staying home also increases to 3 since, using either J-conditionalization or L-conditionalization, $\kappa^{+'}(B) = 3$. Next, suppose that someone tells us that he has a strong hunch that Bill plans to show up for the party, but fails to tell us why. There are two ways in which this report can influence our beliefs. The natural way would be to assume that our informer has not seen Mary's dress and even might not be aware of Bill and Mary's relationship — hence we assess the impact of his report in isolation and say that whatever the value of our current belief in Bill going, it should increase by 3 increments, or $L = 3$. Following Eq. 4.25, $\kappa^{+''}(B)$ and $\kappa^{+''}(\neg M)$ will both be equal to 0,

and we are back to the initial uncertainty about Bill or Mary going to the party, except that our disbelief in both Mary and Bill being at the party has decreased to $\kappa^{+''}(M \wedge B) = 2$. A second way would be to assume that our informer is omniscient and already has taken into consideration all we know about Bill and Mary. He means for us to revise our rankings so that the final belief in "Bill going" will be fixed at $\kappa^{+''}(\neg B) = 3$. With this interpretation, we J-condition $\kappa^{+'}$ on the proposition $\phi = \neg B$ and obtain $\kappa^{+''}(M) = 3$, concluding that it is Mary who will not show up to the party after all.

4.5.3 Complexity Analysis

From Eq. 4.18 we see that $\kappa'(\psi)$ can be computed from $\kappa(\psi|\phi)$ and $\kappa(\psi|\neg\phi)$, which, assuming we have Z^+, requires a logarithmic number of propositional satisfiability tests (see Sec. 4.3). L-conditionalization can follow a similar route (see Eq. 4.25).

Special precautions must be taken when simultaneous, multiple pieces of evidence become available. First, J-conditionalization is not commutative, hence we cannot simply compute κ' by J-conditioning on ϕ_1 and then J-conditioning κ' on ϕ_2 to get κ''. We must J-condition simultaneously on ϕ_1 and ϕ_2 with their respective J-levels, say J_1 and J_2. Worse yet, an auxiliary effort must be expended to compute the J-level of each combination of ϕ's, in our case $\phi_1 \wedge \phi_2$, $\phi_1 \wedge \neg\phi_2$, etc. This is no doubt a hopeless computation when the number of observations is large.

L-conditionalization, by virtue of its commutativity, enjoys the benefits of recursive computations. Let e_1 and e_2 be two (undisclosed) pieces of evidence supporting ϕ_1 (with strength L_1) and ϕ_2 (with strength L_2), respectively. We first L-condition κ on ϕ_1 and calculate $\kappa'(\phi_1)$ and $\kappa'(\phi_2)$ using Eq. 4.24 and Eq. 4.25, respectively. Applying Eq. 4.25 this time to $\kappa'(\psi \wedge \phi_2)$, we calculate $\kappa'(\psi|\phi_2)$. Second, we L-condition κ' on ϕ_2, compute $\kappa''(\phi_2)$ using Eq. 4.24, and finally, using $\kappa'(\psi|\phi_2)$ and $\kappa''(\phi_2)$ in Eq. 4.25 obtain $\kappa''(\psi)$.[22] Note that, although each of these calculations requires only $O(\log|\Delta|)$ satisfiability tests, this computation is effective only when we have a well designated target hypothesis σ to estimate. The computation must be repeated each time we change the target hypothesis, even when the context remains unaltered. This is because we no longer have a facility for economically encoding a complete description of κ', as we had for κ (using the Z^+-function). Thus, the encoding for κ' may not be as *economical* as that for κ (the number of worlds is astronomical), unless

---

[22] The generalization to more than two pieces of evidence is straightforward.

we manage to find dummy rules that emulate the constraints imposed on ϕ_1 by the (undisclosed) observation. Such dummy rules must enforce the conditional independence constraints embedded in Eq. 4.17, without violating the admissibility constraints (Eq. 4.6) in Δ. These dummy rules can be encoded using the stratification mechanism proposed in Chapter 5 (see also [54]).

4.6 Relation to the AGM Theory of Belief Revision

Alchourrón, Gärdenfors, and Makinson (AGM) have advanced a set of postulates that have become a standard against which proposals for belief revision are tested [3]. The AGM postulates model epistemic states as deductively closed sets of (believed) sentences and characterize how a rational agent should change its epistemic states when new beliefs are added, subtracted, or changed. The central result is that the postulates are equivalent to the existence of a complete preordering of all propositions according to their degree of *epistemic entrenchment* such that belief revisions always retain more entrenched propositions in preference to less entrenched ones. Although the AGM postulates do not provide a calculus with which one can realize the revision process or even specify the content of an epistemic state [14, 27, 92], they nevertheless imply that a rational revision must behave as though propositions were ordered on some scale.

Spohn [120] has shown how belief revision conforming to the AGM postulates can be embodied in the context of ranking functions. Once we specify a single ranking function κ on possible worlds, we can associate the set of beliefs with those propositions β for which $\kappa(\neg\beta) > 0$. It follows, then, that the models for the theory ψ representing our beliefs (written $Mods(\psi)$) consist of those worlds ω for which $\kappa(\omega) = 0$. To incorporate a new belief ϕ, one can raise the κ of all models of $\neg\phi$ relative to those of ϕ, until $k(\neg\phi)$ becomes (at least) 1, at which point the newly shifted ranking defines a new set of beliefs. This process of belief revision, which Spohn named α-conditioning (with $\alpha = 1$ for this particular case), was shown to comply with the AGM postulates [33]. It follows then that the process of revising beliefs in all three forms of conditioning also obeys the AGM postulates: Ordinary conditioning amounts to setting $\alpha = \infty$, J-conditioning amounts to $\alpha = J$, while L-conditioning calls for shifting the models of ϕ relative to those of $\neg\phi$ by L units of surprise. If we denote by $\kappa_\phi(\omega)$ the revised ranking after conditioning (with $\alpha = \infty$), then the dynamics of belief is governed by the

following equation:

$$
\kappa_\phi(\omega) = \begin{cases} \kappa(\omega) - \kappa(\phi) & \text{if } \omega \models \phi, \\ \infty & \text{otherwise.} \end{cases}
\tag{4.27}
$$

Accordingly, testing whether a given sentence σ is believed after revision amounts to testing whether $\kappa_\phi(\neg\sigma) > 0$ or, equivalently, whether $\kappa(\neg\sigma|\phi) > 0$.

The unique feature of the system described in this chapter is that the above test can be performed by purely syntactic means, involving only the rules in Δ. These computations are demonstrated in the following example, where the rankings in Tables. 4.3–4.5 are shown for illustrative purposes only.

Example 4.3 (Working students) The set $\Delta = \{s \rightarrow \neg w, s \rightarrow a, a \rightarrow w\}$ stands for "typically students don't work", "typically students are adults", and "typically adults work", respectively.[23] The Z^+-ordering on the rules (computed according to Eq. 4.13) are: $Z^+(a \rightarrow w) = 0$ and $Z^+(s \rightarrow \neg w) = Z^+(s \rightarrow a) = 1$, from which the initial κ^+ ranking can be computed (Eq. 4.7), as depicted in Table 4.3. The rankings in Tables 4.4 and 4.5 show the revised rankings after

κ^+	Possible worlds
0	$(\neg s, a, w), (\neg s, \neg a, w,), (\neg s, \neg a, \neg w,)$
1	$(\neg s, a, \neg w), (s, a, \neg w,)$
2	$(s, a, w), (s, \neg a, \neg w), (s, \neg a, w)$

Table 4.3: Initial ranking for the student triangle in Example 4.3.

observing an adult (κ_a) and a student (κ_s), respectively.

The beliefs associated with these rankings can be computed from the worlds residing in $\kappa^+ = 0$. Thus, in κ_a^+ "an adult works", whereas in κ_s^+ "a student is an adult that does not work". These beliefs can be computed more conveniently by syntactic analysis of the rules and their Z^+-ordering, either by using Eq. 4.14, or by extracting from Δ a propositional theory that is maximally consistent with the observation using Eq. 4.15. For example, the beliefs associated with observing

---

[23]Note that all δ_i's are 0 for this example.

κ_a^+	Possible worlds
0	$(\neg s, a, w)$
1	$(\neg s, a, \neg w), (s, a, \neg w,)$
2	(s, a, w)

Table 4.4: Revised ranking after observing an adult.

κ_s^+	Possible worlds
0	$(s, a, \neg w,)$
1	$(s, a, w), (s, \neg a, \neg w), (s, \neg a, w)$

Table 4.5: Revised ranking after observing a student.

a student s are given by the theory $\{s, s \supset a, s \supset \neg w\}$. These two implications mirror the rules $s \to \neg w$ and $s \to a$ which are the unique set of rules that are maximally consistent with s.

There are several computational and epistemological advantages to basing the revision process on a finite set of conditional rules, rather than on the beliefs or on the rankings or the expectations that emanate from those rules. The number of propositions in one's belief set is astronomical, as is the number of worlds, while the number of rules is usually manageable.

This computational necessity has been recognized by several researchers. For example, Nebel [92] adapted the AGM theory so that finite sets of *base* propositions mediate revisions. The basic idea in this syntax-based system is to define a (total) priority order on the set of base propositions and to select revisions to be maximally consistent relative to that order, as exemplified in the nonmonotonic systems of Brewka [16] and Poole [105] and in Example 4.3. Nebel has shown that such a strategy can satisfy almost all the AGM postulates. Boutilier [14] has further shown that, indeed, the priority function Z^+ corresponds naturally to the epistemic entrenchment ordering of the AGM theory.[24]

---

[24]The proof in [14] considers the priorities Z^+ resulting from a *flat* set of rules as in system-Z [100]. Boutilier [15] also shows that an entrenchment ordering obeying the AGM framework

Unfortunately, even Nebel's theory does not completely succeed in formalizing the practice of belief revision, as it does not specify how the priority order on the base propositions is to be determined. Although one can imagine, in principle, that the knowledge encoder specify this priority order in advance, such specification would be impractical, since the order might (and, as we have seen, should) change whenever new rules are added to the knowledge base. By contrast, system-Z^+ extracts both beliefs and rankings of beliefs automatically from the content of Δ; no outside specification of belief orderings is required.

Finally, and perhaps most significantly, system-Z^+ is capable of responding not merely to empirical observations but also to linguistically transmitted information such as conditional sentences (i.e., if–then rules). For example, suppose someone tells us that "typically, if a person works, that person is compensated" ($w \rightarrow c$); we add this new rule to our knowledge base (verifying first that the addition is admissible), recompute Z^+, and are prepared to respond to new observations or hearsay. In Spohn's system, where revisions begin with a given ranking function κ, one cannot properly revise beliefs in response to new conditional sentences, because, to maintain consistency and coherence, such revision must depend not only on the initial ranking but also on the conditional rules that brought about that initial ranking. Two knowledge bases Δ_1 and Δ_2 might give rise to the same ranking function κ^+ and, yet, the new conditional can be consistent with Δ_1 and inconsistent with Δ_2. As an example, consider the sets $\Delta_1 = \{a \rightarrow b\}$ and $\Delta_2 = \{a \rightarrow b, \neg b \rightarrow \neg a\}$. The ranking κ^+ for these knowledge bases is the same (see Table 4.6). The knowledge base $\Delta_1' = \Delta_1 \cup \{\neg b \rightarrow a\}$ is consistent, as shown on the right-hand side of Table 4.6. On the other hand, the knowledge base $\Delta_2' = \Delta_2 \cup \{\neg b \rightarrow a\}$ is inconsistent. Clearly, these two situations require different procedures for absorbing the new conditional.

κ^+	Δ_1, Δ_2	$\Delta_1' = \Delta_1 \cup \{\neg b \rightarrow a\}$
0	$(a,b), (\neg a, b), (\neg a, \neg b)$	$(a,b), (\neg a, b)$
1	$(a, \neg b,)$	$(a, \neg b,)$
2	Empty	$(\neg a, \neg b,)$

Table 4.6: Ranking κ^+ for $\Delta_1 = \{a \rightarrow b\}$, $\Delta_2 = \{a \rightarrow b, \neg b \rightarrow \neg a\}$, and Δ_1'.

The AGM postulates, likewise, are inadequate for characterizing the process of

obtains from the Z-priorities of the negation of the material counterparts of rules.

incorporating new conditionals, because they are formulated as transformations on belief sets and are thus oblivious to the set of conditionals that shaped those belief sets, and into which the new conditional is about to join.[25]

The ability to adopt new conditionals (as rules) also provides a simple semantics for interpreting nested conditionals (e.g., "if you wear a helmet whenever you ride a motorcycle, then you won't get hurt badly if you fall"[26]). Nested conditionals cease to be a mystery once we permit explicit references to default rules. The sentence "If $(a \rightarrow b)$ then $(c \rightarrow d)$" is interpreted as

> "If I add the default $a \rightarrow b$ to Δ, then the conditional $c \rightarrow d$ will be satisfied by the consequence relation $\vdash_{\div}$ of the resulting knowledge base $\Delta' = \Delta \cup \{a \rightarrow b\}$".

which is clearly a proposition that can be tested in the language of default-based ranking systems. Note the essential distinction between having a conditional sentence $a \rightarrow b$ explicitly in Δ versus having a conditional sentence $a \rightarrow b$ *satisfied* by the consequence relation $\vdash_{\div}$ of Δ. In both cases the conditional $a \rightarrow b$ would meet the Ramsey test, but only the former case would resist the adoption of the conditional $a \rightarrow \neg b$. This distinction gets lost in systems that do not acknowledge defaults as the basis for ranking and beliefs.[27]

4.7   Discussion

This chapter proposes a belief-revision system that reasons semi-tractably and plausibly with linguistic quantification of both observational reports (e.g., "looks like") and domain rules (e.g., "typically"). The system is semi-tractable in the sense that it is tractable for every sublanguage in which propositional satisfiability is polynomial (Horn expressions, network theories, acyclic expressions, etc.). To the best of my knowledge, this is the first system that reasons with approximate probabilities which offers such broad guarantees of tractability. Whereas most tractability results exploit the topological structure of the knowledge base [20, 71, 97] (hypertrees, or partial hypertrees), ours are topology-independent. These results should carry over to the theory of possibility as formulated by Dubois and Prade [28], which has similar features to Spohn's system except that beliefs

---

[25]Gärdenfors [33, pp. 156–160] attempts to devise postulates for conditional sentences, but finds them incompatible with the Ramsey test.

[26]Judea Pearl attributes this example to Philip Calabrese (personal communication).

[27]Belief revision systems proposed in the database literature [31, 19] suffer from the same shortcoming. In that context, defaults represent integrity constraints with exceptions.

are measured on the real interval $[0,1]$. In addition, as Section 4.5 shows, the system can also accommodate expressions of imprecise observations without loss of tractability, thus providing a good model for weighing the impact of evidence and counter-evidence on our beliefs. Also the enterprise of belief revision, as formulated in the work presented in [3, 33], can find a tractable and natural embodiment in system-Z^+, unhindered by difficulties that plagued earlier systems.

From the perspective of defeasible reasoning, system-Z^+ provides the user with the power to explicitly set priorities among default rules, and simultaneously maintains a proper account for specificity relations. However, it inherits some of the deficiencies of system-Z [100][28] the main one being the inability to sanction inheritance across exceptional subclasses (see Exm. 3.2). To illustrate this problem consider adding a a fifth rule $b \overset{\delta_5}{\to} l$ ("birds have legs") to the set of rules in Example 4.1:

r_1: $b \overset{\delta_1}{\to} f$

r_2: $p \overset{\delta_2}{\to} b$

r_3: $p \overset{\delta_3}{\to} \neg f$

r_4: $f \overset{\delta_4}{\to} a$

r_5: $b \overset{\delta_5}{\to} l$

We would normally conclude from this set that "penguins have legs", while system-Z (with $\delta_i = 0$) will consider "penguins" exceptional "birds" with respect to *all* properties, including "having legs". The κ^+-ranking now allows the rule author to partially bypass this obstacle by adjusting the δ's. If δ_5 is set to be bigger than δ_1 (to express perhaps the intuition that anatomic properties are more typical than developmental facilities) then the system will conclude that "typically penguins have legs".[29] This solution however, is not entirely satisfactory. If we add to this new set of rules a class of "birds" which are "legless", system-Z^+ will conclude that either "penguins have legs" or "legless birds fly" but not both.[30] In order to overcome this difficulty, a system must comply with the preference condition in Proposition 3.16. As shown in Section 3.4, maximum

---

[28]And the rational closure described in [72].

[29]Note that the fact that "penguins" are only exceptional with respect "flying" (and not necessarily with respect to "having legs") is automatically encoded in the Z^+ ranking by forcing $Z^+(r_3)$ to exceed $Z^+(r_1) + \delta_3$ independently of δ_5 (and $Z^+(r_5)$).

[30]This counterexample is due to Kurt Konolige.

entropy is one such system; two other systems that satisfy this result are Geffner's Conditional Entailment [36, 38], and the proposal by Boutilier [15].

In Geffner's conditional entailment, rather than letting rule priorities dictate a ranking function on models, a partial order on interpretations is induced instead. To determine the preference between ω and ω', we examine the highest priority rules that distinguish between the two, i.e., that are falsified by one and not by the other. If all such rules remain unfalsified in one of the two possible worlds, then this model is the preferred one. Formally, if $\mathcal{F}[\omega]$ and $\mathcal{F}[\omega']$ stand for the set of rules falsified by ω and ω' respectively, then ω is preferred to ω' iff $\mathcal{F}[\omega] \neq \mathcal{F}[\omega']$ and for every rule in $\mathcal{F}[\omega] - \mathcal{F}[\omega']$ there exists a rule r' in $\mathcal{F}[\omega'] - \mathcal{F}[\omega]$ such that r' has a higher priority than r (written $r' > r$). Thus, a model ω will always be preferred to ω' if it falsifies a proper subset of the rules falsified by ω' (see Prop. 3.16).

Priorities among rules in Geffner proposal differ also from both the proposals in Chapter 3 and this chapter, in that the rule priority relation is a partial order as well. This partial order is determined by the following interpretation of the rule $\varphi \to \psi$: If φ is all we know, then, regardless of other rules that Δ may contain, we are authorized to assert ψ. This means that $r : \varphi \to \psi$ should get a higher priority than any argument (a chain of rules) leading from φ to $\neg\psi$ and, more generally, if a set $\Delta' \subset \Delta$ does not tolerate r, then at least one rule in Δ' ought to have a lower priority than r. In the example above,[31] $r_3 : p \to \neg f$ is not tolerated by the set $\{r_2 : p \to b, r_1 : b \to f\}$, hence we must have that $r_2 < r_3$ or $r_2 < r_1$. Similarly, the rule $r_2 : p \to b$ is not tolerated by $\{r_3 : p \to \neg f, r_1 : b \to f\}$ and hence we also have $r_2 < r_1$ or $r_3 < r_1$. This two conditions together with the transitive properties of $<$, yield $r_2 < r_1$ and $r_3 < r_1$. Note that in this partial order r_4 cannot be compared to any of the other rules. In general, we say that a proposition σ is conditionally entailed by ϕ (in the context of a set Δ) if σ holds in all the preferred models for ϕ induced by every priority ordering *admissible* for Δ. Conditional entailment rectifies many of the shortcommings of system-Z, as well as some weaknesses of the entailment relation induced by maximum entropy. However, having been based on model minimization as well as on enumeration of subsets of rules, its computational complexity might be overbearing. A proof theory for conditional entailment can be found in [36].

Boutilier [15] proposed a system which combines the priority ordering of system-Z (i.e. the flat version of system-Z^+), with Brewka's [16] notion of preferred subtheories. Thus, whereas system-Z^+ assigns equal rank to any two worlds that violate a rule r with $Z^+(r) = z$ and no rule of higher Z^+, the pro-

---

[31]Assuming a flat version where all δ's are zero.

posal in [15] will make further comparisons in terms of rules of lower priority violated in these worlds. In the case above, since any *minimal* world satisfying $p \wedge l$ must violate a proper subset of the rules violated by any *minimal* model for $p \wedge \neg l$, the desired conclusion is certified. These notion are formalized in terms of the modal logic CO^* which is semantically related to the probabilistic interpretation proposed in this dissertation [14]. Nevertheless, counterintuitive examples to this notion of entailment can be found in [36, 48]. While Boutilier's proposal appears to be simpler than conditional entailment (as it does not require partial orders), its computational effectiveness is yet to be analyzed.

CHAPTER 5

Causality

5.1 Introduction

Independently of whether causality is a property of nature or a conceptual convenience, the organization of knowledge as cause–effect relations is fundamental for tasks of prediction and explanation. This chapter introduces, within the basic framework of ranking systems, a simple mechanism called *stratification* for the representation of causal relationships, actions, and changes.

The lack of a mechanism for distinguishing causal relationships from other kinds of associations has been a serious deficiency in most nonmonotonic systems [96], the classical illustration of which is given by the now-famous Yale Shooting Problem (YSP) [57]. In its simplified version, the YSP builds the expectation that if a gun is loaded at time t_0 and Fred is shot with the gun at time t_1, Fred should be dead at time t_2, despite the normal tendency of *being alive* to persist. Many formulations — including circumscription [88], default logic [108], maximum entropy (Chap. 3), system-Z^+ (Chap. 4), and conditional entailment [36] — do not yield the expected conclusion. Instead they reveal an alternative, perfectly symmetrical version of reality, whereby somehow the gun got unloaded and Fred is alive at time t_2.

The inclination to choose the scenario in which Fred dies is grounded in notions of directionality and asymmetry that are particular to causal relationships. This chapter shows that these notions can be derived from one fundamental principle, *Markov shielding*, which can be embodied naturally in preferential model semantics using the device of stratified rankings. Informally, the principle can be stated as follows:

- Knowing the set of causes for a given effect renders the effect independent of all prior events.

In the YSP, given the state of the gun at time t_1, the effect of the shooting can be predicted with total disregard for the gun's previous history.

This chapter proposes a probabilistically motivated, ranked-model semantics

for rules of the form "typically, if $cause_1$ and ... and $cause_n$, then effect", which incorporates the above principle under the assumption that "causes" precede their "effects". As a by-product, our semantics exhibits another feature characteristic of causal organizations: *modularity.* Informally,

- Adding rules that predict future events cannot invalidate beliefs concerning previous events.

This is analogous to a phenomena we normally associate with causal mechanisms such as logical gates in electrical circuits, where connecting the inputs of a new gate to an existing circuit does not alter the circuit's behavior [21].

Although several remedies were proposed for the YSP within conventional nonmonotonic formalisms [118, 36, 121, 8, 79], the formalism explored in this chapter seeks to uncover remedies systematically from basic probabilistic principles [97, pp. 509–516]. Incorporating such principles in the qualitative context of world ranking yields useful results on several frontiers. In prediction tasks (such as the YSP), our formalism prunes the undesirable scenarios, without the strong commitment displayed by *chronological minimization* [118] and without the addition of *external* causal operators to the conditional interpretation of the rules [36] (see Section 5.3). In abduction tasks (such as when Fred is seen alive at t_2), our formalism yields plausible explanations for the facts observed (e.g., similar to [121], the gun must have been unloaded sometime before the shooting at t_1). These suggests that the principle of Markov shielding, by being grounded in probability theory (hence in empirical reality), can provide a coherent framework for the many facets of causation found in commonsense reasoning. Moreover, given the connection formed among causation, defaults, and probability, we can now ask not merely how to reason with a given set of causal assertions but also whether those assertions are compatible with a given stream of observations. A framework for explanations is further discussed in Section 5.3.2.

Section 5.3.1 defines a notion of consistency in the context of causal rules, and briefly compares it to the notion of p-consistency introduced in Chapter 2. Section 5.4 demonstrates how rank-based systems can embody and unify the theories of belief revision [3] and belief updating [65]. Whereas belief revision deals with new information obtained through new observations in a static world, belief update deals with tracing changes in an evolving world, such as that subjected to the external influence of actions.

As shown in Section 4.6 system-Z^+ offers a natural embodiment of the principles of belief revision as formulated by Alchourrón, Gärdenfors and Makinson (AGM) [3], with the additional features of enabling the absorption of new con-

ditional sentences and the verification of counterfactual sentences and nested conditionals. The addition of stratification to system-Z^+, by virtue of representing actions and causation, also provides the necessary machinery for embodying belief updates consistent with the principles proposed by Katsuno and Mendelzon (KM) [65].

5.2 Stratified Rankings

Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a finite set of atomic propositions. Let $c_1, \ldots, c_m$ and e be literals over the elements of $\mathcal{X}$. A *rule* in this chapter is defined as the default $c_1 \wedge \ldots \wedge c_m \rightarrow e$,[1] where the conjunction "$c_1 \wedge \ldots \wedge c_m$" is called the antecedent of the rule and "e" its consequent.[2]

Given $\mathcal{X}$ and a set Δ of rules, the *underlying characteristic graph* for $\langle \mathcal{X}, \Delta \rangle$, is the directed graph $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$ such that there is a node v_i for each $x_i \in \mathcal{X}$, and there is a directed edge from v_i to v_j iff there is a rule r in Δ where x_i (or $\neg x_i$) is part of the antecedent of r, and x_j (or $\neg x_j$) is the consequent of R. We say that Δ is a *causal network* (or *network* for short) if $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$ is acyclic (i.e., $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$ is a DAG). If $v_r, \ldots, v_s$ are the parents of v_t in $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$, then the set $\{x_r, \ldots, x_s\}$ is called the *parent set* of x_t and the set $\{x_r, \ldots, x_s\} \cup \{x_t\}$ is called a *family*. Intuitively, the parent set of an event e represents all the known causes for e. A network Δ induces a strict partial order "$\prec$" on the elements of $\mathcal{X}$ where $x_i \prec x_j$ iff there is a directed path from v_i to v_j in $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$. We will use $\mathcal{O}(\mathcal{X})$ to denote any total order on the elements of $\mathcal{X}$ satisfying $\prec$.[3] Intuitively, $\prec$ represents a natural order on events where causes precede their effects. As an example, Figure 5.1 depicts the underlying graph for the following set of rules to be used in Example 5.1:

r_1: $tk \rightarrow cs$ ("typically, if I turn the ignition key the car starts").

r_2: $tk \wedge bd \rightarrow \neg cs$ ("typically, if I turn the ignition key and the battery is dead, the car will not start").

---

[1] For simplicity we will not introduce a new connective, e.g. $\rightarrow_c$, and we will only consider *flat* causal rules. Section 5.3.3 explores the use of variable strength rules in a causal context.

[2] The form $c_1 \wedge \ldots \wedge c_m \rightarrow e$ does not restrict the development of this chapter but it clarifies the exposition. A causal rule may take on the general form $\alpha(c_1, \ldots, c_m) \rightarrow \beta(e_1, \ldots, e_n)$ where α and β are any Boolean formulae. Any $\alpha(c_1, \ldots, c_m)$ can be simulated by a set of simpler rules, each containing a conjunction of atomic antecedents. Moreover, any rule $\alpha(c_1, \ldots, c_m) \rightarrow \beta(e_1, \ldots, e_n)$ can be represented by the following set of rules: $\alpha(c_1, \ldots, c_m) \rightarrow e'$, $\beta(e_1, \ldots, e_n) \Rightarrow e'$, and $\neg\beta(e_1, \ldots, e_n) \Rightarrow \neg e'$, where e' is a dummy variable and $\Rightarrow$ is a *strict conditional*. The role of strict conditionals in a causal setting is introduced in Section 5.3.

[3] Note that, in particular, any ordering $\mathcal{O}(\mathcal{X})$ induced by a topological sort on the nodes of $\Gamma_{\langle \mathcal{X}\Delta \rangle}$, where $x_i < x_j$ if v_i precedes v_j in the topological sort, satisfies $\prec$.
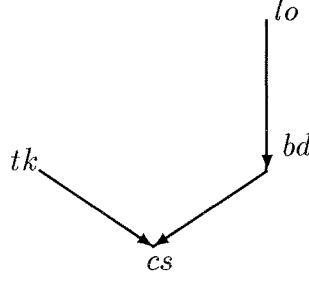
Figure 5.1: Underlying graph for the causal rules in the battery example

r_3: $lo \rightarrow bd$ ("typically, if I leave the headlights on all night the battery is dead").

In previous chapters, the interpretation of a rule $\varphi \rightarrow \psi$ was based on the condition of admissibility of κ (see Def. 4.2). A ranking κ is admissible relative to Δ iff for every $\varphi_i \rightarrow \psi_i \in \Delta$:[4]

$$\kappa(\neg\psi_i|\varphi_i) > 0 \tag{5.1}$$

We now extend this requirement and introduce a stratification constraint that will endow the rules with a causal character.

Definition 5.1 (Stratified Rankings) Given a network Δ, an admissible ranking κ relative to Δ, and an ordering $\mathcal{O}(\mathcal{X})$; let X_i ($1 \leq i \leq n$) denote a literal variable taking values from $\{x_i, \neg x_i\}$, and let Par_{X_i} denote the conjunction $X_r \wedge \ldots \wedge X_s$ where $\{X_r, \ldots, X_s\}$ is the parent set of x_i. We say that κ is *stratified* for Δ under $\mathcal{O}(\mathcal{X})$, if for $2 \leq i \leq n$, and for any instantiation of the variables $X_1, \ldots, X_i$, we have

$$\kappa(X_i|X_{i-1} \wedge \ldots \wedge X_1) = \kappa(X_i|Par_{X_i}) \tag{5.2}$$

$\square$

Eq. 5.2 says that in a stratified ranking the incremental surprise of finding x_i in a full description of some past scenario, must be equal to the incremental surprise of finding x_i given just the state of Par_{X_i} in that same scenario. Thus, the parent set of an event x_i (Par_{X_i}) *shields* this event x_i from all prior events (see Fig. 5.2). This condition parallels the Markovian independence conditions

---

[4]Assuming all δ_i's are equal to zero; otherwise admissibility would require that $\kappa(\neg\psi_i|\varphi_i) > \delta_i$ in Eq.5.1.
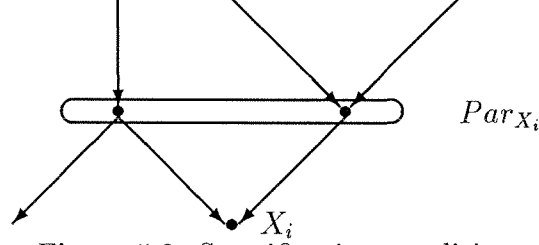
Figure 5.2: Stratification condition.

embodied in Bayes Networks (BN) [97].

A BN is a pair $\langle \Gamma, P \rangle$ where Γ is a DAG and P is a probability distribution. Each node v_i in Γ corresponds to a variable X_i in P, and P decomposes into the product:

$$P(X_n, \ldots, X_1) = \prod_{i=1}^{i=n} P(X_i | Par_{X_i}) \tag{5.3}$$

which, similarly to Eq. 5.2, incorporates the assumption that the parent set of any given variable X_i renders X_i probabilistically independent of all its predecessors (in the given ordering). Causal networks can in fact be regarded as an order-of-magnitude abstraction of BN's, where exact numerical probabilities are replaced by integer-valued levels of surprise (κ), addition is replaced by min, and multiplication is replaced by addition (see [53, 120, 102]). Eq. 5.2 can be re-written to mirror Eq. 5.3 as:[5]

$$\kappa(X_n \wedge \ldots \wedge X_1) = \sum_{i=1}^{i=n} \kappa(X_i | Par_{X_i}) \tag{5.4}$$

Note that Eq. 5.4 also constitutes an effective test for checking whether a given ranking κ is stratified for an arbitrary network Δ. The test can be made recursive if we write Eq. 5.4 as

$$\kappa(X_m \wedge \ldots \wedge X1) = \sum_{i}^{m} k(X_i | Par_{X_i}), \ \text{m=1,2,\ldots,n} \tag{5.5}$$

which follows from Eq. 5.4 after *marginalizing*[6] over $\{Xn, \ldots, Xm + 1\}$, $m = 1, 2, \ldots, n$. We shall show that the requirement of stratification augments ad-

---

[5]An even coarser abstraction of Eq. 5.3 in the context of relational databases can be found in [21], where the stratification condition is imposed on relations and then used in finding backtrack free solutions for constraint satisfaction problems.

[6]In probability theory, we marginalize over $\{X_n, \ldots, X_{m+1}\}$ by summing over all instantiations for these variables; thus, we have $P(X_m, \ldots, X_1) = \sum_{X_n, \ldots, X_{m+1}} P(X_n, \ldots, X_1)$. It follows from Eqs. 4.1–4.3 that $\kappa(X_m \wedge \ldots \wedge X_1) = \sum_{X_n, \ldots, X_{m+1}} \kappa(X_n \wedge \ldots \wedge X_1)$.

missible rankings with the properties of Markov shielding and modularity (see Theorems 5.6 and 5.7 below), that we normally attribute to causal organizations.

The following theorem states that the stratification criterion (Eq. 5.2) does not depend on the specific ordering $\mathcal{O}(\mathcal{X})$. This implies that in order to test whether a given ranking κ is stratified relative to a network Δ, it is enough to test Eq. 5.2 against *any* ordering $\mathcal{O}(\mathcal{X})$.

Theorem 5.2 *Given a network Δ, let $\mathcal{O}_1(\mathcal{X})$ and $\mathcal{O}_2(\mathcal{X})$ be two orderings of the elements in $\mathcal{X}$ according to Δ. If κ is stratified for Δ under $\mathcal{O}_1(\mathcal{X})$, then κ is stratified for Δ under $\mathcal{O}_2(\mathcal{X})$.*

To illustrate the nature of stratification, we will compare two admissible rankings associate with the network $\Delta = \{a \to \neg c, b \to c\}$. A stratified ranking for Δ is depicted on the left-hand side of Table 5.1 (κ_s), while the ranking on the right-hand side represents the κ^+ (system-Z^+) ranking for Δ.[7] In order to show

κ	κ_s: Stratified	κ^+: System-Z^+
0	$(\neg a, b, c)$, $(\neg a, \neg b, c)$, $(\neg a, \neg b, \neg c)$	$(\neg a, b, c)$, $(\neg a, \neg b, c)$, $(\neg a, \neg b, \neg c)$, $(a, \neg b, \neg c)$
1	$(a, \neg b, \neg c)$, $(a, b, \neg c)$, $(\neg a, b, \neg c)$	(a, b, c), $(a, b, \neg c)$, $(\neg a, b, \neg c)$, $(a, \neg b, c)$
2	(a, b, c), $(a, \neg b, c)$	no worlds in this rank

Table 5.1: Stratified, κ^*, and κ^+ rankings for $\{a \to \neg c, b \to c\}$.

that κ'^+ is not stratified we select the order $\mathcal{O} = (A, B, C)$ (which agrees with the characteristic DAG of Δ) and test whether $\kappa^+(\neg c \wedge b \wedge a)$ satisfies Eq. 5.4. From Table 5.1 $\kappa^+(\neg c \wedge a \wedge b) = 1$, $\kappa^+(\neg c | a \wedge b) = \kappa^+(a) = \kappa^+(b) = 0$, and therefore $\kappa^+(\neg c \wedge a \wedge b) \neq \kappa^+(\neg c | a \wedge b) + \kappa^+(a) + \kappa^+(b)$ contrary to the requirements of Eq. 5.4. Alternatively, we can use the Markov shielding property (to be proven in Thm. 5.6) according to which the parents of every variable render that variable independent of all its other predecessors. Since B is a root node in the characteristic DAG of Δ, it has no parents, and it must therefore be (marginally) independent of all its predecessors, namely of A. In terms of ranking functions this requirement of independence translates into

$$\kappa(A \wedge B) = \kappa(A) + \kappa(B) \qquad (5.6)$$

---

[7]The maximum entropy ranking κ^* for this network Δ is identical to κ^+.

for all instantiations of the literals A and B (taking values from $\{a, \neg a\}$ and $\{b, \neg b\}$ respectively). As can be verified from Table 5.1, κ_s complies with Eq. 5.6.[8]

5.3   c-Entailment

Given a network Δ each stratified ranking κ defines a consequence relation $\Vdash_{\kappa}$ where $\phi \Vdash_{\kappa} \sigma$ iff $\kappa(\sigma \wedge \phi) < \kappa(\neg\sigma \wedge \phi)$ or if $\kappa(\phi) = \infty$. A consequence relation is said to be *proper* for $\phi \Vdash_{\kappa} \sigma$ iff $\kappa(\phi) \neq \infty$.

Definition 5.3 (c-Entailment) A network Δ *c-entails* σ given ϕ, written $\phi \Vdash_{\Delta} \sigma$, iff $\phi \Vdash_{\kappa} \sigma$ in every κ stratified for Δ, which is proper for $\phi \Vdash_{\kappa} \sigma$.
□

In other words, given Δ, we can expect σ from the evidence ϕ, iff the preference constraint conveyed by $\phi \to \sigma$ is satisfied by every stratified ranking for Δ. Def. 5.3 parallels the definition of $\vdash_{p}$ (probabilistic entailment, Def. 3.9) with the only difference being that the rankings for $\Vdash_{\Delta}$ must be stratified. We remark that c-entailment is not to be interpreted as stating that ϕ is believed to *cause* σ. Rather, it expresses an *expectation* to find σ true in the context of ϕ, having given a causal character to the rules in Δ.

Since the set of stratified rankings for a given Δ is a subset of the admissible rankings for Δ, every stratified consequence relation must satisfy the rules of inference of *Logic, Cummulativity* and *Cases* introduced in Section 3.2. It follows then, that the rules of inference below are sound for c-entailment.

Theorem 5.4 *Let* $\varphi, \psi, \gamma, \sigma$, *and their conjunction be satisfiable wffs. The following are sound rules of inference for* $\Vdash_{\Delta}$:

1. **(Defaults)** *If* $\varphi \to \psi$ *(or* $\varphi \Rightarrow \psi$*)* $\in \Delta$ *then* $\varphi \Vdash_{\Delta} \psi$.

2. **(Logic)** *If* $\models \varphi \supset \psi$ *then* $\varphi \Vdash_{\Delta} \psi$.

3. **(Augmentation)** *If* $\varphi \Vdash_{\Delta} \psi$ *and* $\varphi \Vdash_{\Delta} \gamma$ *then* $\varphi \wedge \gamma \Vdash_{\Delta} \psi$.

4. **(Cut)** *If* $\varphi \Vdash_{\Delta} \gamma$ *and* $\varphi \wedge \gamma \Vdash_{\Delta} \psi$ *then* $\varphi \Vdash_{\Delta} \psi$.

5. **(Cases)** *If* $\varphi \Vdash_{\Delta} \psi$ *and* $\gamma \Vdash_{\Delta} \psi$ *then* $\varphi \vee \gamma \Vdash_{\Delta} \psi$.

The first rule (*Default*) follows immediately from the requirement of admissibility. Rules 2 and 5 correspond to the rules of *Logic* and *Cases* of Section 3.2, and rules 3

---

[8]Eq. 5.6 can be also obtained from 5.5 by setting $m = 2$.

and 4 simple rewrite the *Cumulativity* rule of Section 3.2 by breaking the *iff* into two cases.

The following are derived rules of inference and further illustrate the logical properties of c-entailment and will be used in Examples. 5.1 and 5.2.[9]

Theorem 5.5 *Derived rules of inference:*

1. **(Deductive closure)** *If $\varphi \mathrel{\Vdash_\Delta} \psi$ and $\varphi \mathrel{\Vdash_\Delta} \gamma$ and $\models \varphi \wedge \psi \wedge \gamma \supset \sigma$ then $\varphi \mathrel{\Vdash_\Delta} \sigma$*

2. **(Presuppositions)** *If $\varphi \mathrel{\Vdash_\Delta} \psi$ and $\varphi \wedge \gamma \mathrel{\Vdash_\Delta} \neg\psi$ then $\varphi \mathrel{\Vdash_\Delta} \neg\gamma$.*

3. **(And)** *If $\varphi \mathrel{\Vdash_\Delta} \psi$ and $\varphi \mathrel{\Vdash_\Delta} \gamma$ then $\varphi \mathrel{\Vdash_\Delta} \psi \wedge \gamma$.*

These rules of inference are also sound with respect to p-entailment (and probabilistic entailment) and, therefore, as discussed in previous chapters, are too weak to constitute a full account of plausible reasoning. The next two theorems provide additional inference power (reflecting the stratification condition) which emanates from the causal structure of Δ. They establish conditions under which these inference rules can be applied modularly to subsets $\Delta' \subset \Delta$ with the guarantee that the resulting inferences will hold in Δ.

Theorem 5.6 *Let Δ be a network, and let $\{p_r, \ldots, p_s\}$ be a set of literals corresponding to the parent set $\{x_r, \ldots, x_s\}$ of x_t (each p_i, $r \leq i \leq s$, is either x_i or $\neg x_i$). Let e_{x_t} denote a literal built on x_t, and let $\mathcal{Y} = \{y_1, \ldots, y_m\}$ be a set of atomic propositions such that no $y_i \in \mathcal{Y}$ is a descendant of x_t in $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$. Let ϕ_Y be any wff built only with elements from $\mathcal{Y}$ such that $\phi_Y \wedge p_r \wedge \ldots \wedge p_s$ is satisfiable. If $p_r \wedge \ldots \wedge p_s \mathrel{\Vdash_\Delta} e_{x_t}$ then $\phi_Y \wedge p_r \wedge \ldots \wedge p_s \mathrel{\Vdash_\Delta} e_{x_t}$.*

Theorem 5.7 *Let $\mathcal{X}' \subset \mathcal{X}$ and $\Delta' \subset \Delta$ such that all rules in Δ' are built with atomic propositions in $\mathcal{X}'$, and if $x' \in \mathcal{X}'$ then all the rules in Δ with either x' or $\neg x'$ as their consequent are also in Δ'. Let φ and ψ be two wffs built with elements from $\mathcal{X}'$. If $\varphi \mathrel{\Vdash_{\Delta'}} \psi$ then $\varphi \mathrel{\Vdash_\Delta} \psi$.*

These theorems confirm that stratified rankings exhibit the properties of Markov shielding and modularity. As a corollary to Theorem 5.7 it is easy to see that c-entailment is insensitive to *irrelevant* propositions, and moreover, given two networks with no causal interaction, their respective sets of plausible conclusions will

---

[9]They are taken from [36] where a formal derivation in terms of the rules of inference in Theorem 5.4 can be found.

be independent of each other. To obtain a complete proof theory for c-entailment the four axioms of graphoids [97, Chapter 3] need to be invoked.[10] Theorems 5.6 and 5.7 cover the essence of these axioms and are sufficiently powerful to illustrate the main features of c-entailment. Consider the following example:[11]

Example 5.1 (Dead battery) The network $\Delta = \{tk \rightarrow cs, tk \wedge bd \rightarrow \neg cs, lo \rightarrow bd\}$ encodes the information that "typically if I turn the ignition key the car starts", "typically if I turn the ignition key and the battery is dead the car will not start", and "typically if I leave the head lights on all night the battery is dead". The underlying graph for this network is depicted in Figure 5.1. Given Δ, and the fact the we left the head lights on all night, we don't expect the car engine to start once we turn the ignition key (i.e., $lo \wedge tk \not\vdash_{\Delta} \neg cs$). As in the case of YSP, an unintended scenario exists, in which the car engine actually starts and the battery is not dead after all. In both maximum entropy and system-Z^+, for example, $\kappa^+(lo \wedge tk \wedge cs) = \kappa^+(lo \wedge tk \wedge \neg cs)$ and $\kappa^*(lo \wedge tk \wedge cs) = \kappa^*(lo \wedge tk \wedge \neg cs)$, and consequently neither $lo \wedge tk \not\vdash_+ \neg cs$ nor $lo \wedge tk \not\vdash_* \neg cs$. The reason for this behavior is that the $\kappa(\omega)$ in these approaches depends on the priorities of rules violated in ω and the priorities assigned to rules do not properly reflect their relative position in the causal structure. Given that the key is turned and the lights were left on, we know that either the rule $tk \rightarrow cs$ or the rule $lo \rightarrow bd$ must be violated.[12] In both these approaches, these rules receive the same priority, and therefore the unintended scenario is as *normal* as the intended one.[13]

Table 5.2 contains an example of a stratified ranking for Δ, showing the inequality $\kappa(lo \wedge tk \wedge \neg cs) < \kappa(lo \wedge tk \wedge cs)$. Note that the surprise $\kappa = 1$ associated with the world $\omega = lo \wedge bd \wedge tk \wedge \neg cs$ is not caused by any rule violation, but rather, by the abnormality of event lo (as well as bd), whose κ is indeed 1. Although the rule $tk \rightarrow cs$ is violated in ω, it does not contribute any additional surprise to $\kappa(\omega)$ over and above $\kappa(lo)$. Note also that the abnormality of the event lo was not explicitly indicated by the rule author. Rather, it is was deduced from the stratified structure of Δ which must render lo and tk independent, hence, if lo is abnormal when tk is true (because one of the two rules must be violated) it must

---

[10]The conditional independence defined by $\kappa(X_3|X_2, X_1) = \kappa(X_3|X_2)$ is clearly a graphoid since κ represents infinitesimal probabilities (See [120, 62]).

[11]This example is isomorphic to the YSP [36].

[12]A third possibility is that $tk \wedge bd \rightarrow \neg cs$ is violated; but since this is a more specific rule than $tk \rightarrow cs$ its Z-priority will be higher (in both maximum entropy and system-Z^+, and therefore no minimal model for either $lo \wedge tk \wedge cs$ or $lo \wedge tk \wedge \neg cs$ will violate this rule.

[13]We could force the desired conclusion by setting the strengths δ of the rules to the appropriate values. This, however, would require advanced knowledge of all the rules in the knowledge base (and their interactions). The objective in this chapter is a formalism able to extract the necessary information automatically.

also be abnormal when tk is false – refraining from turning the key cannot make us believe that the lights were left on or that the battery is dead. This ability to infer that lo (as well as bd) is an abnormal eventuality, a rather compelling inference intuitively, is what distinguishes stratified ranking from maximum entropy and system-Z^+.[14] Proposition 5.8 presents a formal derivation of $lo \wedge tk \Vdash_{\Delta} \neg cs$:

κ	worlds
0	$(\neg lo, \neg bd, tk, cs)$, $(\neg lo, \neg bd, \neg tk, \neg cs)$
1	$(lo, bd, tk, \neg cs)$, $(lo, bd, \neg tk, \neg cs)$, $(\neg lo, bd, tk, \neg cs)$, $(\neg lo, bd, \neg tk, \neg cs)$
2	$(lo, \neg bd, tk, cs)$, $(lo, \neg bd, \neg tk, \neg cs)$
3	Rest of the ω's

Table 5.2: Stratified ranking for $\{tk \rightarrow cs, tk \wedge bd \rightarrow \neg cs, lo \rightarrow bd\}$.

Proposition 5.8 $lo \wedge tk \Vdash_{\Delta} \neg cs$

Proof: Let $\mathcal{X}' = \{lo, bd, tk\}$ and let $\Delta' = \{lo \rightarrow bd\}$.

1. $lo \Vdash_{\Delta'} bd$; by the *Defaults* rule.

2. $tk \wedge lo \Vdash_{\Delta'} bd$; by 1 and Theorem 5.6.

3. $tk \wedge lo \Vdash_{\Delta} bd$; by 2 and Theorem 5.7.

4. $tk \wedge bd \Vdash_{\Delta} \neg cs$; by the *Defaults* rule.

5. $tk \wedge bd \wedge lo \Vdash_{\Delta} \neg cs$; by 4 and Theorem 5.6.

6. $tk \wedge lo \Vdash_{\Delta} \neg cs$; by 3, 5 and the *Cut* rule.

$\square$

The key intermediate steps in this derivation rely on Theorems 5.6 and 5.7, which embody the principles of markov shielding and modularity:

---

[14]If the rule $True \rightarrow \neg bd$ is added to Δ, system-Z^+ would yield the expected conclusion.

- $tk \wedge lo \,\|\!\!\sim_{\Delta} bd$. This follows from the proposition tk and applying Theorem 5.7 to the sub-network Δ' built from the language $\mathcal{X}' = \{lo.bd.tk\}$ and containing only the rule $lo \rightarrow bd$.

- $tk \wedge bd \,\|\!\!\sim_{\Delta} \neg cs$ and $tk \wedge bd \wedge lo \,\|\!\!\sim_{\Delta} \neg cs$. The former follows directly from p-entailment, and the latter from applying Theorem 5.6 to the rule $tk \wedge bd \rightarrow \neg cs$, and the proposition lo.

The next example presents a simple abduction (or backward projection) problem, and permits us to compare the behavior of c-entailment with that of chronological minimization [118].

Example 5.2 (Unloading the gun.) Consider

$$\Delta = \{l_0 \rightarrow l_1, l_1 \rightarrow l_2, \ldots, l_{n-1} \rightarrow l_n\}$$

standing for the various instances of "typically, if a gun is loaded at time t_i, then it is expected to remain loaded at time t_{i+1}" $(0 \leq i < n)$. We say that a rule $l_i \rightarrow l_{i+1}$ is *falsified* by ω iff $\omega \models l_i \wedge \neg l_{i+1}$; a stratified ranking κ relative to Δ can be constructed as follows:

$$\kappa(\omega) = \text{number of rules in } \Delta \text{ falsified by } \omega \qquad (5.7)$$

Given that the gun is loaded at t_0 and that it is found unloaded at time t_n (i.e., $l_0 \wedge \neg l_n$ is true), the scheme of chronological minimization will favor the somewhat counterintuitive inference that the gun remained loaded until t_{n-1} (i.e., $l_1 \wedge \ldots \wedge l_{n-1}$ is true). c-entailment on the other hand, only yields the weaker conclusion that the gun must have been unloaded any time within t_1 and t_{n-1} (i.e., $\neg(l_1 \wedge \ldots \wedge l_n)$), but the exact instant where the "unloading" of the gun occurs remains uncertain.

Proposition 5.9 $l_0 \wedge \neg l_n \,\|\!\!\sim_{\Delta} \neg(l_1 \wedge \ldots \wedge l_n)$

Proof: Follows trivially from the *Deduction* rule. The fact that we cannot point out the exact moment in which the gun is unloaded follows from the ranking built by Eq. 5.7, since all formulas representing these situations have equal ranking. $\square$

c-entailment and chronological minimization are expected to yield the same conclusions in problems of pure prediction, since enforcing ignorance of future events is paramount to the principle of modularity, which was shown to be inherent to c-entailment. They differ however in tasks of abduction, as demonstrated

in Example 5.2. In this respect, c-entailment is closer to both *motivated action theory* [121] and *causal entailment* [36]. However, contrary to the motivated action theory, c-entailment automatically enforces specificity-based preferences, which are natural consequences of the conditional interpretation of rules.[15]

We end this section by discussing the *strict* version of a causal rule denoted by $\Rightarrow$, which will be useful in representing non-defeasible causal influences in Sections 5.3.2 and 5.4. Semantically, strict rules impose the following constraints on the admissibility conditions of a ranking κ (Eq. 5.1): for each $\varphi \Rightarrow \psi$ in the knowledge base,

$$\kappa(\psi \wedge \varphi) < \kappa(\neg\psi \wedge \varphi) = \infty, \quad \text{and} \quad \kappa(\varphi) < \infty. \tag{5.8}$$

Intuitively, a strict conditional voids interpretations that render its antecedent true and its consequent false by assigning them the lowest possible preference; a rank κ equal to infinity.[16] The following are two properties of strict rules:

Proposition 5.10 *Let* $C_1 \wedge \ldots \wedge C_n \Rightarrow E \in \Delta$

1. **(Contraposition)** *If there exists a stratified ranking for* Δ *where* $\kappa(\neg E) < \infty$ *then* $\neg E \ \|\!\sim_\Delta \neg(C_1 \wedge \ldots \wedge C_n)$

2. **(Transitivity)** *If* $\varphi \ \|\!\sim_\Delta \psi$ *and* $\psi \models (C_1 \wedge \ldots \wedge C_n)$ *then* $\varphi \ \|\!\sim_\Delta E$

These properties mirror the behavior of the material implication "$\supset$", but the resemblance is in fact only superficial. As discussed in Sections 2.1 and 2.7, the semantic difference between a strict rule $c \Rightarrow e$ and the wff $c \supset e$ is that the former expresses necessary hence permanent constraints while the latter expresses information bound to the current situation. Thus, the former participates in constraining the admissible rankings while the latter is treated as an "observation" formula $\neg c \vee e$, and can affect conclusions only by entering the antecedents of queries. This difference is greatly accentuated when strict conditionals are treated as causal rules, because stratified rankings are more sensitive to the rule format. Indeed, contraposing the rule $a \Rightarrow b$ into $\neg b \Rightarrow \neg a$ changes the causal relationship between a and b and this change should reflect on the resulting rank. Compare, for example, $\Delta_1 = \{c \rightarrow \neg b, a \Rightarrow b\}$ and $\Delta_2 = c \rightarrow \neg b, \neg b \Rightarrow \neg a$. Any stratified ranking for Δ_1 must render a and c totally independent of each other, as two unrelated causes of the variable B.

---

[15]We remark that the formalism in [121] deals with a much richer time ontology than the formalism presented here, and with a first-order language.

[16]This is equivalent to requiring that $P(\psi|\varphi) = 1$ and $P(\varphi) > 0$ (see 2.2).

5.3.1 c-Consistency

Chapter 2 proposed a norm of consistency, called p-consistency for rules conveying prototypical information. This norm and its associated decision procedure (Sec. 2.4) were shown to be sufficient when rules were augmented with degrees of strength (Thm. 4.3). The semantical requirements of stratification induce a new notion of consistency, specific to the causal interpretation of rules, which is radically different from p-consistency.

Definition 5.11 (c-Consistency) A network Δ is *c-consistent* iff there exists at least one stratified ranking κ for Δ.
$\square$

An example of a c-inconsistent network is $\Delta = \{tk \rightarrow cs, tk \wedge bd \rightarrow \neg cs, tk \rightarrow x, x \rightarrow bd\}$.[17] To show that Δ is inconsistent note that by *Presupposition* (Thm. 5.5) we have $tk \not\Vdash_{\Delta} \neg bd$, which implies that in all stratified rankings $\kappa(\neg bd \wedge tk) < \kappa(bd \wedge tk)$. A simple application of Theorem 5.6 on $x \rightarrow bd$ (and the proposition tk) yields $tk \wedge x \Vdash_{\Delta} bd$. By the *Defaults* rule $tk \Vdash_{\Delta} x$ which together with $tk \wedge x \Vdash_{\Delta} bd$ and the *Cut* rule yields $tk \Vdash_{\Delta} bd$, which in turn implies the contradictory inequality $\kappa(bd \wedge tk) < \kappa(\neg bd \wedge tk)$. The lack of an appropriate causal interpretation for this set of rules is not surprising. If we accept that tk causes cs, we should expect $\neg bd$ to hold by default when tk is true. On the other hand, if there is a *causal path* from tk to bd, we should expect bd to hold in the context of tk. Note that this set is p-consistent.

We can find an admissible but not stratified ranking for Δ (see Table 5.3).[18] This ranking depicts a situation in which the act of predicting the consequences of turning the key seems to protect the battery against the damage inflicted by x, and such a flow of events is indeed contrary to the common understanding of causation. In fact, if we do not ascribe a causal character to the rules, we cannot apply Theorem 5.6 and thus $tk \Vdash_{K} bd$ is not in the consequence relation of all admissible rankings.

Another c-inconsistent set is $\Delta = \{a \Rightarrow c, b \Rightarrow \neg c\}$, which might arise when we physically connect the outputs of two logic gates with conflicting functions. Since neither a nor b have parents in $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$, every stratified ranking (for Δ) must yield

$$\kappa(a \wedge b) = \kappa(a) + \kappa(b), \tag{5.9}$$

---

[17]This is the network used in Example 5.1 augmented with the two rules $tk \rightarrow x$ and $x \rightarrow bd$.

[18]This ranking is not stratified for Δ since $\kappa(bd \wedge x \wedge tk) = 2$, but $\kappa(bd|x) + \kappa(x|tk) + \kappa(tk) = 1$, which contradicts Eqs. 5.2 and 5.4.

κ	worlds
0	$(\neg tk, x, bd, \neg cs)$
1	$(tk, x, \neg bd, cs)$
2	$(tk, x, bd, \neg cs)$
3	Rest of the ω's

Table 5.3: Admissible ranking for $\{tk \rightarrow cs, tk \wedge bd \rightarrow \neg cs, tk \rightarrow x, x \rightarrow bd\}$.

implying that a and b are independent events. However, if each time we observe a we should expect c and, each time we observe b we should expect $\neg c$, then a and b must be mutually exclusive, hence negatively correlated events. Indeed, since $\kappa(a \wedge b \wedge c) = \kappa(a \wedge b \wedge \neg c) = \infty$, we have $\kappa(a \wedge b) = \infty$, and Eq. 5.9 cannot be satisfied unless either a or b is permanently false, thus defying the "possible antecedent" requirement for strict rules (Eq. 5.8). Note that this Δ is again p-consistent since, if it were not for the requirement of Eq. 5.9, an admissible ranking can be constructed by simply excluding (by setting $\kappa = \infty$) any ω such that $\omega \models a \wedge b$, which would still permit us to assign $\kappa(a) = \kappa(b) < \infty$.

5.3.2 Accountability: A Framework For Explanations

Causality is a worthy abstraction of complex interactions in as much as it proves itself useful for formulating predictions and explanations for modeling these interactions. The bulk of the effort in previous sections was spent in incorporating into ranking representations properties associated with causality and showing how these properties can be used to facilitate prediction. In this section we concentrate on producing plausible explanations for a given set of observations.

For example, once we are told that "turning the ignition key *causes* the car engine to start" we would like to *explain* a car-engine running by conjecturing that somebody must have turned the ignition key. However, $cs \Vdash_\Delta tk$ is not a c-entailed conclusion from the network $\Delta = \{tk \rightarrow cs\}$. The problem is that we haven't provided any information in Δ that establishes the starting of the car as a phenomena that in itself needs to be explained. The rule $tk \rightarrow cs$ only imposes two constraints on the rankings of possible worlds: First, cs should hold in all the most preferred models for tk and, second, once tk is known to be true we can expect cs to hold independently of any event prior to tk. But this says little about

97

the models of $\neg cs$ and, in particular, whether the car will start without turning the ignition key. In normal discourse, when given a set of rules "C_i causes E" we usually subscribe (by convention) to additional assumptions that help complete this information.[19] Three of the most common assumptions are:

- **Accountability:** An effect E is presumed *false* if all the conditions listed as causes of E are also *false*.[20]

- **Exception Independence:** The rules representing the "cause-effect" relations may admit exceptions which inhibit the occurrence of the effect even in prescence of the cause. However, unless explicitly stated or logically implied, these exceptions are presumed independent. In the car example, a dead-battery and an empty gas tank can be considered as such exceptions to $tk \rightarrow cs$. Both will prevent the car engine from starting, and are presumed to be independent of each other.

- **Disjunctive Interaction:** The likelihood of an event does not diminish when several of its causes prevail simultaneously. For example, if *rain* and *sprinkler-on* are each a cause for the grass being wet, the grass will be only *more* likely to be wet if both the sprinkler is turned on and it is raining.[21]

A probabilistic model that captures these assumptions, called *noisy-or gate*, is described in [97] where it is proposed as a canonical model of disjunctive interaction among causes $C_1, \ldots, C_n$ that predict the same effect E. The noisy-or gate is depicted schematically in Fig. 5.3. The set $\{I_1, \ldots, I_n\}$ represents inhibitors, where each I_i stands for an abnormality that would interfere with the causal connection between C_i and E. Every pair C_i and I_i constitutes the inputs to an and-gate so that if C_i is "active" (or true) and I_i is not known to be active, then the output s_i will provide *support* for the effect E. Each s_i is then an input to the final or-gate. If one or more of these s_i's is active then E is expected to be true, and if all s_i are false, then E is expected to be false.

Both and-gates and or-gates impose functional constraints on propositions; thus, in order to represent their behavior strict rules are necessary. The rules in Eqs. 5.10–5.13 formalize the intended behavior of an and-gate:

$$C_i \wedge \neg I_i \Rightarrow s_i \tag{5.10}$$

---

[19]See [68] for a discussion on the relation between completing the information and abduction reasoning for tasks of producing explanations from observations.

[20]This may require that we lump together all unknown causes of E under the heading "all other causes".

[21]This assumption actually follows from that of exception independence.
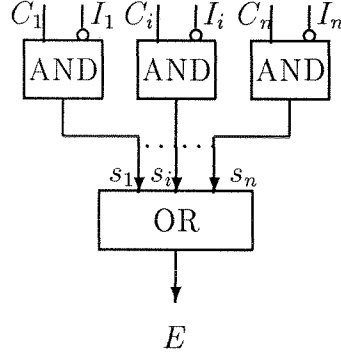
Figure 5.3: The *noisy-or* interaction model: $C_1, \ldots, C_n$ are the set of causes for E, and each I_i represents an *inhibitor* or *abnormality* for $C_i \to E$

"If C_i is true and I_i is not active, then there is support s_i for E",[22]

$$\neg C_i \;\Rightarrow\; \neg s_i \tag{5.11}$$

"if the cause C_i is not active there is no support s_i for E."

$$I_i \Rightarrow \neg s_i. \tag{5.12}$$

"If I_i is active there is no support s_i for E."

$$True \to \neg I_i \tag{5.13}$$

"I_i is an abnormality, so it is false by default".

The **or-gate** represents the interaction between the set of causal rules for the effect E, with propositions $s_1, \ldots, s_n$ as inputs and the literal constant E as output. The behavior of this gate is governed by a pair of strict rules: [23]

$$s_1 \vee \ldots \vee s_n \Rightarrow E, \tag{5.14}$$

and a *closure* rule

$$\neg(s_1 \vee \ldots \vee s_n) \Rightarrow \neg E \tag{5.15}$$

This last rule incorporates the assumption of accountability: If there is no causal support for E, then E must be false. Strict rules are necessary to simulate the disjunctive nature of the or-gate since $s \wedge s' \not\mid\!\sim_\Delta e$ is not c-entailed from $\Delta = \{s \vee s' \to e\}$.

---

[22]This rule is reminiscent of the proposed encoding of defaults under circumscription using the *ab* predicate suggested by McCarthy [88].

[23]Note that we could have equivalently encoded Eq. 5.14 as a set of rules $s_i \Rightarrow E$, $1 \le i \le n$, since n applications of the *Disjunction* rule of inference (Thm 5.4) on $s_i \Rightarrow E$ will in fact yield $s_1 \vee \ldots \vee s_n \Rightarrow E$.
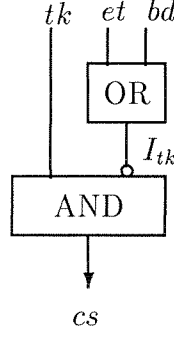
Figure 5.4: A Schematic view of the car example.

Proposition 5.12 *Given a network Δ such that $s_1 \vee \ldots \vee s_n \Rightarrow E \in \Delta$, then $(\bigwedge_{i=j}^{i=k} s_i) \mid\!\!\!\sim_\Delta E$, where $1 \leq j \leq k \leq n$.*

Thus, given a set of causal relations "C_i causes E", $1 \leq i \leq n$, we can use the rules in Eqs. 5.14 and 5.15 for modeling the **or-gate** and the rules in Eqs.5.10–5.13 for modeling the **and-gates**.

In many cases we have explicit knowledge of the identity of the mechanisms capable of inhibiting the normal causal connection between C_i and E. Their interactions can be modeled in the same fashion, using **and-gates** and **or-gates** as building blocks. For example, given that "turning the key (tk) causes the car engine to start (cs)" and two mechanisms that might inhibit this relation, namely a dead battery (db) and an empty gas tank (et), we would require a noisy-or for the $tk \to cs$ causal relation:[24]

$$tk \wedge \neg I_{tk} \Rightarrow cs \quad ; \quad True \to \neg I_{tk} \tag{5.16}$$

$$I_{tk} \Rightarrow \neg cs \quad ; \quad \neg tk \Rightarrow \neg cs \tag{5.17}$$

and another noisy-or to model the interaction between the two causes bd and et for the inhibitor I_{tk}. Lets assume, for simplicity, that these causal relations are strict and void of any inhibitors themselves. Thus, we can simplify this noisy-or to a *standard* **or-gate**:

$$bd \vee et \Rightarrow I_{tk} \quad ; \quad \neg(bd \vee et) \Rightarrow \neg I_{tk} \tag{5.18}$$

Fig 5.4 presents a schematic view of this example and Table 5.4 contains a stratified ranking. Some c-consequences of the rules in Eqs. 5.16–5.18 are:

$$tk \mid\!\!\!\sim_\Delta cs \quad ; \quad cs \mid\!\!\!\sim_\Delta tk \quad ; \quad tk \wedge bd \mid\!\!\!\sim_\Delta \neg cs \tag{5.19}$$

---

[24]Since in this case there is only one cause for cs, we simplify the encoding and skip the final or-gate.

100

κ	worlds ω
0	$(tk, \neg I, \neg bd, \neg et, cs)$, $(\neg tk, \neg I, \neg bd, \neg et, \neg cs)$
1	$(tk, I, bd, \neg et, \neg cs)$, $(tk, I, \neg bd, et, \neg cs)$, $(\neg tk, I, \neg bd, et, \neg cs)$, $(\neg tk, I, bd, \neg et, \neg cs)$
2	$(tk, I, bd, et, \neg cs)$, $(\neg tk, I, bd, et, \neg cs)$,

Table 5.4: A stratified ranking for the car example.

Given the independence constraints embedded in our formalism, any stratified ranking for the rules in Eqs. 5.16–5.18 will comply with $\kappa(bd \wedge et) = \kappa(bd) + \kappa(et)$ making a world where both bd and et are true more abnormal than one in which only one of them holds. Thus, in the situation in which we turn the key and the car engine does not start, c-entailment conclude that either the battery is dead or the gas-tank is empty, but not both:

$$tk \wedge \neg cs \mathrel{\|\!\!\!\sim_{\Delta}} ((bd \wedge \neg et) \vee (\neg bd \wedge et)). \tag{5.20}$$

In Section 5.3.3 we explore mechanisms to add degrees of strength to the rules using the formalism described in [53], so that the degrees of support for each hypothesis can be used to manipulate the focus of the diagnosis process. To complete the encoding of the causal relations in Example 5.1 we add an **and-gate** representing the causal relation between "head lights on all night" (*lo*) and the dead battery (*bd*):

$$lo \wedge \neg I_{lo} \Rightarrow bd \quad ; \quad True \rightarrow \neg I_{lo} \tag{5.21}$$

$$I_{lo} \Rightarrow \neg bd \quad ; \quad \neg lo \Rightarrow \neg bd \tag{5.22}$$

This proposal of model completion requires that the set of causes for a given effect be both identifiable and separable from the set of causes that prevent the effect, i.e. the inhibitors. One way of establishing this difference is by eliciting the information directly from the rule encoder: For each effect E we would ask for a list of causes $C_i, \ldots C_n$ and a list of events (causes or effects) that might prevent E from occurring. Another way is to allow the input of the causal relations to be specified in the same language of networks. Then the system would *compile* this network into a target network containing rules in Eqs. 5.16–5.18 filling in the assumptions of the noisy-or model using **and-gates** and **or-gates** as building

blocks. This process would examine the rules *family* by *family*[25] following the stratified order imposed by the underlying graph of Δ. In case of *conflicting* relations, i.e. a set of literal supporting an effect e and another set supporting $\neg e$, the system will try to uncover the inhibitor from the causes by using the rule of *Presupposition* in Theorem 5.5. This rule of inference says that if $C_i \mathrel{|\!\!\!\sim_\Delta} E$ and $C_i \wedge I_i \mathrel{|\!\!\!\sim_\Delta} \neg E$ then $C_i \mathrel{|\!\!\!\sim_\Delta} \neg I_i$. Thus, in the car example we would have the set $\Delta = \{tk \rightarrow cs, bd \wedge tk \Rightarrow \neg cs, et \wedge tk \Rightarrow \neg cs\}$ as input, and a simple application of the rule of *Presupposition* will mark both bd and et as inhibitors with respect to tk. Note that cases of *ambiguous* families like $\{a \rightarrow c, b \rightarrow \neg c\}$ would require further information about the relation between a and b, since first we cannot distinguish between causes and inhibitors, and second, an encoding of both these rules as **and-gates** will result in a c-inconsistent network similar to the example of conflicting strict arrows in Section 5.3.1. The problem is that the assumption of exception independence is no longer valid: The cause for c (i.e., a) is the inhibitor for the cause of $\neg c$ (i.e., b) and vice-versa.

5.3.3 The most normal stratified ranking

In Section 5.3.2 we saw that in order to reap the full benefits of Bayesian Networks, we needed to supplement the constraints of Δ with additional information that further shapes the conditional rankings of each family in the underlying DAG. Another approach of supplementing the missing information is to establish a preference relation among stratified rankings and rule out those rankings that are less preferred than others. Since a lower ranking is associated with greater *normality*, it is natural that out of the set of all admissible stratified rankings we prefer those that assign to interpretations the lowest possible ranks, and then define the entailment relation with respect to this set of privileged rankings.

Such a strategy was adopted in Chapter 4 (without the requirement of stratification) and led to system-Z^+. The incorporation of the most-normal strategy in the context of stratified rankings, will result in a substantial increase of expressiveness. For example, in the noisy-or encoding of causal relations, assuming independence among the inhibitors, the most-normal (minimal) stratified ranking κ_c^+ is given by the following function:

$$\kappa_c^+(\omega) = \text{number of inhibitors that are true in } \omega \tag{5.23}$$

Thus, in the car example (with dead battery and empty tank as inhibitors) in Eqs. 5.16–5.18, the minimal ranking $\kappa_c^+(\omega)$ will be 0, 1 or 2 depending on whether

---

[25]A family is the set of propositions composed by the parent set of an effect and the effect itself (Section 5.2).

$\omega \models \neg bd \wedge \neg et$, $\omega \models (bd \wedge \neg et) \vee (\neg bd \wedge et)$ or $\omega \models bd \wedge et$ respectively (this minimal ranking is depicted in Table 5.4).

The incorporation of variable-strength rules in this context is especially useful: From knowing that we turned the key but the car did not start, it follows that either the battery is dead or that the tank is empty ($tk \wedge \neg cs \mathrel{\vert\!\sim_{\widehat{\Delta}}} ((bd \wedge \neg et) \vee (\neg bd \wedge et))$). However, if we knew that an empty tank is more likely than a dead battery, we could encode this information as $True \overset{\delta_1}{\rightarrow} \neg et$ and $True \overset{\delta_2}{\rightarrow} \neg bd$ with $\delta_1 < \delta_2$. The minimal ranking $\kappa_c^+(\omega)$ in this case will be 0, $\delta_1 + 1$, $\delta_2 + 1$, or $\delta_1 + \delta_2 + 1$ depending on whether $\omega \models \neg bd \wedge \neg et$, $\omega \models (\neg bd \wedge et)$, $\omega \models (bd \wedge \neg et)$ or $\omega \models bd \wedge et$ respectively. Now given the context that we turn the key and the car does not start, our primary suspect would be the lack of gasoline $tk \wedge \neg cs \mathrel{\vert\!\sim_{\widehat{\Delta}}^*} et$, where $\mathrel{\vert\!\sim_{\widehat{\Delta}}^*}$ denotes the consequence relation of the most-normal stratified ranking (which is unique for this example).

Note, as shown in Table 5.5, that the most-normal stratified ranking may not be unique for the network $\Delta = \{a \rightarrow c, b \rightarrow \neg c\}$. Therefore, we need to define entailment in minimal rankings, denoted by $\mathrel{\vert\!\sim_{\widehat{\Delta}}^*}$, with respect to the consequence relations of all most-normal stratified rankings.

κ	Rank 1
0	$(\neg a, b, \neg c), (\neg a, \neg b, c), (\neg a, \neg b, \neg c)$
1	$(\neg a, b, c), (a, b, c), (a, \neg b, c)$
2	$(a, b, \neg c), (a, \neg b, \neg c)$

κ	Rank 2
0	$(a, \neg b, c), (\neg a, \neg b, c), (\neg a, \neg b, \neg c)$
1	$(\neg a, b, \neg c), (a, b, \neg c), (a, \neg b, \neg c)$
2	$(a, b, c), (\neg a, b, c)$

Table 5.5: Two minimal rankings for $\{a \rightarrow c, b \rightarrow \neg c\}$

It is not clear at this point whether this loss of uniqueness will result in substantial increase in computational complexity. Although we may lose the semi-tractability of system-Z^+, we can still exploit the topological properties

of the characteristic DAG ($\Gamma_{\langle \mathcal{X}, \Delta \rangle}$) to render practical reasoning feasible. It is well known that the local propagation techniques of Bayesian Networks can be extended to sparse networks by embedding the network in a hypertree (or acyclic database) [115]. Thus, it is quite feasible that similar techniques could be applied to the computation of consequences in systems governed by the "most-normal" completion of stratified rankings.

5.4 Belief Update

Section 4.6 demonstrated how the semantics of model ranking, together with the syntactic machinery developed for system-Z^+, can be applied to manage the tasks of belief revision, in conformity with the AGM postulates. The introduction of stratified ranking adds the capability for implementing a new type of belief changes, named *update* by Katsuno and Mendelzon (KM) [65]. In both tasks (belief revision and belief update) we seek to incorporate a new piece of information ϕ into an existing set of beliefs ψ. Yet, in belief revision ϕ is assumed to be a piece of evidence while in update ϕ is treated as a change occurring by external intervention. Katsuno and Mendelzon [65] have shown that the AGM postulates are inadequate for describing changes caused by updates, for which they have proposed new sets of postulates. The basic difference between revision and update is that the latter permits changes in each possible world independently, as was proposed by Winslett [127].[26]

Belief update can be embodied in a stratified ranking system using the following device: For each instruction to "update the knowledge base by ϕ" we add a set of rules that simulates the action "$do(\phi)$, leaving everything else constant (whenever possible)", and then condition κ on the truth of $do(\phi)$. The following set of causal rules embody the intent of this action, where ϕ and ϕ' stand for "ϕ holds at t" and "ϕ holds at $t' > t$", respectively:[27]

$$\phi \;\rightarrow\; \phi' \tag{5.24}$$

$$\neg\phi \;\rightarrow\; \neg\phi' \tag{5.25}$$

$$do(\phi) \;\Rightarrow\; \phi'. \tag{5.26}$$

The following example (adapted from Winslett [127]) demonstrates how this device differentiates between update and revision.

---

[26]In the language of Bayesian networks, the difference between updates and revisions parallels the distinction between causal and evidential information [96].

[27]The two persistence rules, Eqs. 5.24 and 5.25, are presumed to apply between any two atomic propositions at two successive times.

Example 5.3 (XOR-gate) A XOR Boolean gate $c = XOR(a, b)$ is examined at two different times. At time t, we observe the output $c = true$ and conclude that one of the inputs a or b must be true, but not both. At a later time t' we learn that b' is true (primed letters denote propositions at time t'), and we wish to change our beliefs (in a and a') accordingly. Naturally, this change should depend on how the truth of b' is learned. If we learn b' by measuring the voltage on the b terminal of the gate, then we have a belief revision process on our hands, and we expect a' to be false. On the other hand, if we learn that b' is true as a result of physically connecting the b terminal to a voltage source, we no longer expect a' to be false, since we have no reason to believe that the output c has retained its truth value in the process.

In the stratified ranking formulation, the knowledge base corresponding to this example will consist of three components:

1. The functional description of the XOR gate at times t and t',

$$a \wedge b \Rightarrow \neg c \quad ; \quad \neg a \wedge \neg b \Rightarrow \neg c \tag{5.27}$$

$$a \wedge \neg b \Rightarrow c \quad ; \quad \neg a \wedge b \Rightarrow c, \tag{5.28}$$

and an equivalent set of rules for a', b', c'.

2. The persistence rules: For every x in $\{a, b, c\}$,

$$x \rightarrow x' \quad ; \quad \neg x \rightarrow \neg x'. \tag{5.29}$$

3. The action $do(b)$, which represents the external influence on b':

$$do(b) \Rightarrow b'. \tag{5.30}$$

The underlying graph for the network Δ corresponding to this knowledge base is depicted in Figure 5.5.

Initially, after observing c, our evidence consists only of c. The minimal stratified ranking κ_c for a Δ consisting of rules in Eqs. 5.27-5.30 is depicted in Table 5.6. To represent belief revision, we add b' to our evidence set and query whether $c \wedge b' \; \|\!\!\sim_\Delta^* \neg a'$.[28] In contrast, to represent belief update, we add $do(b)$ to our evidence set and query whether $(c \wedge b' \wedge do(b)) \; \|\!\!\sim_\Delta^* \neg a'$.

It is easy to show that the first query is answered in the affirmative, while the second in the negative. The left-hand side of Table 5.7 shows the ranking

---

[28]Recall that $\|\!\!\sim_\Delta^*$ denotes the consequence relation of the minimal stratified ranking for Δ (see Sec. 5.3.3), which is unique for this example.
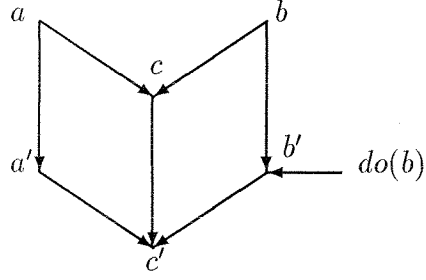
Figure 5.5: Graph depicting the causal dependencies in Example 5.3

κ_c	$\neg do(b)$	$do(b)$
0	$(\neg a, b, \neg a', b'), (a, \neg b, a', \neg b')$	
1	$(\neg a, b, a', b'), (a, \neg b, \neg a', \neg b'), (a, \neg b, a', b')$	$(\neg a, b, \neg a', b'), (a, \neg b, a', b')$
2	$(\neg a, b, a', \neg b'), (a, \neg b, \neg a', b'),$	$(\neg a, b, a', b'), (a, \neg b, \neg a', b')$
∞	models for $\neg c$	models for $\neg c$

Table 5.6: Minimal stratified ranking for Example 4 after c is observed

κ_c	Revision $\kappa_c(\omega | b')$	Update $\kappa_c(\omega | do(b))$
0	$(\neg a, b, \neg a', b')$	$(\neg a, b, \neg a', b'), (a, \neg b, a', b')$
1	$(\neg a, b, a', b'), (a, \neg b, a', b')$	$(\neg a, b, a', b'), (a, \neg b, \neg a', b')$
2	$(a, \neg b, \neg a', b'),$	
∞	models for $\neg b'$	models for $\neg do(b)$

Table 5.7: Rankings after observing b, and after "doing" b

106

resulting from revising the ranking in Table 5.6 by b' (first query), while the right-hand side shows the ranking after updating by $do(b)$ (second query). Note that in the revised ranking the only world in the zero rank is a model for $\neg a'$, while the updated ranking shows an additional world which is a model for a' (the state of the output c in this world changed as a consequence of the action). The action $do(b)$ establishes the truth of b' but has no effect on what we believe about the second input a'. Since neither a nor $\neg a$ were believed at t, they remain unbelieved at t'.

5.4.1 The dynamics of belief update

The example above demonstrates that, given a ranking κ and a network Δ, it is possible to predict how a system would respond to external interventions. For example, if we wish to inquire whether event e will hold true after we force some variable A to become true, we simply add to Δ the rule $do(a) \Rightarrow a$,[29] recompute the resulting stratified ranking κ' on the augmented set of variables (including $do(a)$), and then compute $\kappa'(e|do(a))$. There is a simple relation between $\kappa(e|a)$ and $\kappa'(e|do(a))$, which results in a direct transformation between two ranking functions, $\kappa(\omega)$ and $\kappa_{do(a)}(\omega)$, the latter being an abbreviation of $\kappa'(w|do(a))$. From Eq. 5.4 we have that:

$$\kappa(X_n \wedge \ldots \wedge A \wedge \ldots \wedge X_1) \;=\; \sum_{i=1}^{i=n} \kappa(X_n | Par_{X_n}) \tag{5.31}$$

$$\kappa(X_n \wedge \ldots \wedge A \wedge \ldots \wedge X_1) \;=\; \sum_{i=1, i \neq j}^{i=n} \kappa(X_n | Par_{X_n}) + \kappa(A | Par_A) \tag{5.32}$$

where A is the jth literal taking values from $\{a, \neg a\}$. Similarly, the stratification of κ' relative to $\Delta \cup \{do(a) \Rightarrow a\}$ dictates

$$\kappa'(X_n \wedge \ldots \wedge A \wedge \ldots \wedge X_1 \wedge DO(a)) =$$
$$\sum_{i=1, i \neq j}^{i=n} \kappa'(X_n | Par_{X_n}) + \kappa'(A | Par'_A) + \kappa'(DO(a)) \tag{5.33}$$

---

[29]We use lowercase to denote the instantiation of variable A to a truth value.

Where $DO(a)$ is a variable taking values from $\{do(a), \neg do(a)\}$, $Par'_A = Par_A \cup \{Do(a)\}$, and

$$
\kappa'(A|Par'_A) = \begin{cases} 0 & \text{if } A = a \text{ and } DO(a) = do(a), \\[2mm] \kappa(A|Par_A) & \text{if } DO(a) = do(a), \\[2mm] \infty & \text{if } A = \neg a \text{ and } DO(a) = do(a). \end{cases} \tag{5.34}
$$

Eq. 5.33 reflects the fact that the action variable $DO(a)$ is a root node in $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$, since it is under the sole *control* of the rule author, while Eq. 5.34 reflects the constraint $do(a) \Rightarrow a$. Since the new rule $do(a) \Rightarrow a$ only affects the family of A, we have that the summation term in Eq. 5.33 is equal to the summation term in Eq. 5.32. Conditioning Eq. 5.33 on $do(a)$ and making the appropriate substitutions yields

$$
\kappa'(X_n \wedge \ldots \wedge A \wedge \ldots \wedge X_1 | do(a)) = \kappa(X_n \wedge \ldots \wedge A \wedge \ldots \wedge X_1) -
$$
$$
- \kappa(A|Par_A) + \kappa'(A|Par_A \wedge do(a)) \tag{5.35}
$$

Where, according to Eq. 5.34, $\kappa'(A|Par_A \wedge do(a)) = 0$ when $A = a$, and $\kappa'(A|Par_A \wedge do(a)) = \infty$ when $A = \neg a$. Thus, making again the appropriate substitutions we get

$$
\kappa_{do(a)}(\omega) = \begin{cases} \kappa(w) - \kappa(a|Par_A(w)) & \text{if } \omega \models a. \\[2mm] \infty & \text{if } \omega \models \neg a. \end{cases} \tag{5.36}
$$

In other words, the κ of each world ω satisfying a is reduced by an amount equal to the degree of surprise of finding $A = true$, given the realization of Par_A in ω (denoted by $Par_A(\omega)$. The κ of each world falsifying a is of course ∞.[30]

Such independent movement from world to world is shown in Example 5.3, where $\kappa(\omega)$ is depicted on the left-hand side of Table 5.6 and $\kappa_{do}(\omega)$ is depicted on the right hand side of Table 5.7. If A has no *parents* (direct causes), then $\kappa_{do(a)}$ is obtained by shifting the κ of each $\omega \models a$ by a constant amount $\kappa(a)$, as in ordinary conditioning, and $\kappa_{do(a)}(\omega)$ would be equal to $\kappa(\omega|a)$, as expected. However, when the manipulated variable has direct causes Par_A, the amount of

---

[30]The reader might recognize Eq. 5.36 in its probabilistic form where, given a probability function $P(\omega)$ and a causal network Γ, the probability $P'(\omega)$ obtained by manipulating variable A to take on the value a is given by: $P'(\omega) = P(\omega)/P(a|Par_A(w))$ (for $\omega \models a$). This can be easily shown from the functional definition of causal relationships as used, for example, in Pearl and Verma [103].

shift would vary from world to world, depending on how surprising it would be (in that world) to find a happening naturally (without external intervention). For instance, if A is governed by persistence rules, $a(t-1) \to a(t), \neg a(t-1) \to \neg a(t)$, then worlds in which $a(t-1)$ hold will shift less than those in which $a(t-1)$ is false, because $a(t)$ is expected to hold in the former and not in the latter. Note that the amount of shift subtracted from $\kappa(\omega)$ is equal precisely to the fraction of surprise $\kappa(a|Par_A(\omega))$ that $A = true$ contributes to $\kappa(\omega)$ and that now becomes *explained away* (hence excusable) by the action $do(a)$. The generalization of Eq. 5.36 to the case where a conjunction of literals $\phi = a_1 \wedge \neg a_2 \ldots$ are forced to become true or false is straightforward:

$$\kappa_{do(\phi)}(\omega) = \kappa(w) - [\kappa(a_1|Par_{A_1}(\omega)) + \kappa(\neg a_2|Par_{A_2}(\omega)) + \ldots] \qquad (5.37)$$

if $\omega \models \phi$, and $\kappa_{do(\phi)}(\omega) = \infty$ otherwise.

Note that any stratified ranking κ has at least one variable A possessing a remarkable invariant properties:

$$\kappa_{do(a)}(\omega) = \begin{cases} \kappa(\omega) & \text{if } \omega \models a, \\[2mm] \infty & \text{otherwise.} \end{cases} \qquad (5.38)$$

Intuitively, every variable satisfying Eq. 5.38 corresponds to a sink in $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$, or a "last" variable in the "temporal" ordering $\mathcal{O}$. Indeed, for any such sink, Eq. 5.38 conveys the intuition that by manipulating the last variable in the temporal order, we do not expect the past to change. It is comforting to see that the ramifications of the Markov shielding principle coincide with an alternate reading of causation as a specification of behavior under external interventions.

5.4.2 Relation to KM postulates

Katsuno and Mendelzon [65] have formulated belief update as a transformation between two formulas, ψ, representing our current set of beliefs, and ϕ, the new information we wish to incorporate into that set of beliefs. The update process is assumed to be an operator $\diamond$ that takes the formula ψ and transforms it into a new formula $\psi \diamond \phi$, that syntactically represents our updated set of beliefs. KM have introduced a set of postulates which characterize all update operators that can be defined by the possible world approach of Winslett [127], hence, they are considered universal conditions for any model describing belief change due to external actions. One such postulate, for example,

(U2) If ψ implies ϕ, then $\psi \diamond \phi$ is equivalent to ψ,

says that if the new sentence ϕ is derivable from belief set ψ, that updating by ϕ does not alter the belief set.

In our ranked-based model of beliefs change, the current stock of beliefs is represented by those worlds ω for which κ is zero. Hence, ψ is defined by the union of all ω such that $k(w) = 0$. If the new sentence (ϕ) is a conjunction of literals, then the updated ranking is given by Eq 5.37 and the new set of beliefs, $\psi \diamond \phi$, is represented semantically by the union of all worlds w for which $\kappa_{do(\phi)}(w)$ is zero.[31]

That updates resulting from Eq. 5.36 comply with the KM postulates can be seen by the following consideration. KM have shown that their axioms are equivalent to the existence of a function mapping each possible interpretation world ω to a partial pre-order $\leq_\omega$, such that for any interpretation ω', if $\omega \neq \omega'$ then $\omega <_\omega \omega'$. Then the set of models for the update of a formula ψ (representing our current beliefs) by a formula ϕ, written $\psi \diamond \phi$, is found by taking the union of the minimal models for ϕ, with respect to each one of the pre-orders defined by the models for ψ:

$$Mods(\psi \diamond \phi) = \bigcup_{\omega \in Mods(\psi)} \min(Mods(\phi), \leq_\omega). \tag{5.39}$$

In other words, Eq. 5.39 asserts that the models of $\psi \diamond \phi$ can be obtained by replacing each ψ-world ω with a set of ϕ-worlds ω^* that are *nearest* to ω. We shall call each such w^* an *image* of ω, a word coined by Lewis [77] to denote a counterfactual alternative to Bayes conditioning. If ω is consistent with ϕ then its image ω^* is equal to ω itself, as is required by $\leq_\omega$. However, when ω is inconsistent with ϕ, its image is a closest (according to $\leq_\omega$) world satisfying ϕ.

Thus, to show compliance with the KM postulates we need to define a preorder $\leq_\omega$ and show that for every world $\omega \models \neg\phi$ that is currently assigned $\kappa(\omega) = 0$, Eq. 5.37 takes each image ω^* of ω and moves it toward $\kappa_{do(\phi)}(\omega^*) = 0$. We shall construct such a preorder and show that, moreover, in an image world ω^*, every

---

[31]Updates involving disjunctions require special treatment. If they are to be interpreted as a license to effect any change satisfying the disjunction, then the final state of belief is the union, taken over all disjuncts, of worlds that drift to $\kappa = 0$. In this interpretation, the instruction "make sure the box is painted either blue or white" will leave the box color unknown, even knowing that the box was white initially (contrary to the postulate (U2) of KM). However, if the intention is to effect no change as long as the disjunctive condition is satisfied, then the knowledge base should be augmented with an observation-dependent strategy "$do(\phi)$ when ϕ is not satisfied", instead of using the pure action $do(\phi)$. Conditioning on such a strategy again yields a belief set consistent with the KM postulates. The first interpretation is useful for discrediting earlier observations, for example, "I am not sure the employee's salary is $50K$; it could be anywhere between $40K$ and $60K$".

term $\kappa(x_i|Par_{X_i}(\omega^*)) > 0$ represents a violation of expectation that would be totally excusable were it caused by an external intervention such as ϕ. Intuitively, the image world corresponds to a scenario in which all the unexpected events are attributed to the intervention of ϕ but otherwise the world follows its natural, unperturbed course as dictated by the prediction of the causal theory.

It is not hard to see that the image ω^* as described above is indeed a minimal element in the order $\leq_\omega$, defined as follows:

Definition 5.13 (World orderings) Let $\mathcal{O} = x_1, x_2, \ldots, x_n$ be any order of the variables that is consistent with the DAG $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$. Given three worlds ω, ω_1, and ω_2, we say that $\omega_1 \leq_\omega \omega_2$ iff the following conditions hold:

1. ω disagrees with ω_2 on a literal that is earlier (in $\mathcal{O}$) than any literal on which ω disagrees with ω_1.

2. If a tie occurs, then $w_1 \leq_\omega \omega_2$ if $\kappa(\omega_1) \leq k(w_2)$.

$\square$

Theorem 5.14 *Let ψ be a wff representing a set of beliefs. Let κ be a ranking such that $\omega \in Mods(\psi)$ iff $\kappa(\omega) = 0$. Let ϕ represent a conjunction of literals, and let $\kappa_{do(\phi)}$ be the ranking that results from updating κ by ϕ as shown in Eq. 5.36 such that $\omega^* \in Mods(\psi \diamond \phi)$ iff $\kappa_{do(\phi)}(\omega^*) = 0$. Then*

$$Mods(\psi \diamond \phi) = \bigcup_{\omega \in Mods(\psi)} \min(Mods(\phi), \leq_\omega). \tag{5.40}$$

5.4.3 Related work

The connection between belief update and theories of action was noted by Winslett in [127] and has been elaborated more recently by del Val and Shoham [22] using the situation calculus. In fact, del Val and Shoham [22] showed that the KM-postulates can be derived from their formulation of actions in the situation calculus, as they are derived from the theory presented in this chapter. The interesting power of these postulates is that they cover a wide variety of such formulations, from a simple theory such as the one introduced here to the intricate machinery of the situation calculus. Due to their broad generality, the KM postulates should not be taken as a complete characterization of actions-based updates, but merely as a useful norm of coherence on the resulting belief change. The analysis in this chapter offers the KM postulates an intuitive, model-theoretic support that is

well grounded in probability theory, and is accompanied with a concrete characterization of causation and action. It also offers a simple unification of revision and update, since both are embodied in a conditioning operator, the former by conditioning on *observations* and the latter by conditioning on *actions*.

Grahne et. al. [56] showed that revision could be expressed in terms of an update operator in a language of introspection (intuitively, *observing* a piece of evidence has the same effect as *causing* the observer to augment her beliefs by that very evidence). The analysis in this chapter shows that the converse is also true: belief updates can be expressed in terms of a *conditioning* operator, which is normally reserved for belief revision. The intuition is that *acting* to produce a certain effect yields the same beliefs as *observing* that action performed. This translation is facilitated by the special status that the added *action* $\Rightarrow$ *effect* rules enjoy in stratified ranking, where actions are always represented as root nodes, independent of all other events except their consequences. This ensures that the immediate effects of those actions are explained away and do not reflect back on other events in the past. It is this stratification that produces the desired distinction between observing an action produce an effect and observing the effect without the action.[32]

5.5 Discussion

Extensions to the formalism proposed in this chapter should include efficient decision procedures for c-consistency and c-entailment, and a complete proof theory for c-entailment. Also of interest, are notions of entailment based on strategies for completing the information provided by the rules in a network Δ. In Section 5.3.3 we presented one such strategy based on the most normal completion proposed in Chapter 4. However, contrary to the case of system-Z^+, this strategy will not always yield a unique ranking for the stratified case. Further investigations are needed to uncover classes of networks where the resulting ranking is unique and the decision procedures tractable. The bridge that the principle of Markov shielding establishes between probabilistic and nonmonotonic formalisms invites insights on these issues, including efficient query answering procedures and

---

[32]Note that update cannot be expressed in terms of the AGM operators of revision and contraction, because it is impossible to simulate with these operators the acceptance of a new conditional $do(\phi) \Rightarrow \phi$, so that the acceptance of $do(\phi)$ is treated differently than the acceptance of ϕ. Similarly, update cannot be formulated as a transformation on rankings such as Spohn's conditioning because the identity of the image world ω^* cannot be described in terms of the initial ranking alone; it requires the causal theory Δ. Two different theories, Δ_1 and Δ_2 may give rise to the same ranking κ, and still require two very different updates.

methods of completion (e.g., the noisy-or canonical model in Sec. 5.3.2), from the literature on Bayes networks.

As pointed out in Section 5.3, the notion of c-entailment $\varphi \parallel_{\Delta} \psi$ should not be understood as establishing φ as a cause of ψ. The reason is best illustrated through the following example. Consider the network $\Delta = \{True \rightarrow b\}$; it follows from this network that $a \parallel_{\Delta} b$ where a is an arbitrary proposition in the language. By the condition of stratification, each stratified ranking κ must comply with $\kappa(A \wedge B) = \kappa(A) + \kappa(B)$.[33] Therefore, in every stratified ranking for Δ it must be true that

$$\kappa(a \wedge b) < \kappa(a \wedge \neg b), \text{ or} \tag{5.41}$$

$$\kappa(a) + \kappa(b) < \kappa(a) + \kappa(\neg b) \tag{5.42}$$

since the rule in Δ establishes that $k(b) < \kappa(\neg b)$. Thus, the reason that b is expected given a is not because a is a cause for b, in fact a is actually independent of b. b is expected simply because it is true by default according to Δ. Note that this problem disappears if we make the additional requirement that $\kappa(\neg b|a) > \kappa(\neg b)$. However, this definition of causation, by being based on Bayesian conditioning, would still be subject to the classical difficulties with spurious correlations (or hidden causes). The definition proposed here is based on the use of rules like $do(c) \Rightarrow c$ to simulate an external manipulation of the causes. The decision on whether c is a cause of e can be made based on manipulating and observing the behavior of e. Thus, for example, c can be identified as a cause for e in the context of a knowledge base Δ, if $do(c) \parallel_{\Delta} e$ in the context of $\Delta \cup \{do(c) \Rightarrow c\}$, but it is not the case that $do(\neg c) \parallel_{\Delta} e$ in the context of $\Delta \cup \{do(\neg c) \Rightarrow \neg c\}$. This notion is in line with the counterfactual reading of causation in Lewis [76] where asserting that c is a cause of e implies that e would not have occurred if it were not for c. It is also in line with the control-based reading of causation which underlies most statistical tests for causal influences as well as the method proposed by Pearl and Vermain [103] for discovering causality in nonexperimental studies. This interpretation reads:

> "c is a cause for e if an external agent interfering only with c can affect y."

In nonexperimental studies the external agent is simulated by a "virtual-control" variable, while in our formulation it is enacted by the *do* operator; both must comply with the Markov shielding constraint.

---

[33]Capital letters denote literal variables.

Finally, we point out that the probabilistic roots of the semantics proposed provides bi-directional inferences for causal and evidential information, the potential of refining pre-encoded knowledge by learning from experience, and the usual guarantees of clarity, coherence and plausibility that accompany theories grounded in empirical reality. Some of the other novel contributions of this chapter are: A consistency norm for knowledge bases representing causal relationships, uniform and practical formulations for belief revision, belief updating, and general reasoning about action and change.

CHAPTER 6

Concluding Remarks

6.1   Summary

This dissertation is an account of a semantical and computational approach to reasoning with incomplete and defeasible information encoded as default rules. These rules are regarded as if–then conditional sentences allowing exceptions that have different degrees of abnormality. Semantically, the rules are interpreted using infinitesimal probabilities, which can be viewed as qualitative abstractions of an agent's experience. An equivalent semantics is provided that interprets the rules using ranks on models, where higher ranked models stand for more surprising (or less likely) situations. Computationally, these semantics admit effective procedures for testing the consistency of knowledge bases containing default rules and for computing whether (and to what degree) a given query is confirmed or denied. The result is a model-theoretic account of plausible beliefs that, as in classical logic, are qualitative and deductively closed and, as in probability, are subject to retraction and to varying degrees of firmness.

The probabilistic semantics enable the introduction of principled ways for solving some of the problems with irrelevance that plagued previous conditional based approaches. This is accomplished by restricting the set of rankings that are considered admissible with a given knowledge base. At the heart of this formulation is the concept of *default priorities*, namely, a natural ordering of the conditional sentences that is derived automatically from the knowledge base and is used to answer queries without computing explicit rankings of worlds or formulas. As a result, some query-answering procedures (those for system-Z^+; see Chap. 4) require only a polynomial number of propositional satisfiability tests and hence tractable for Horn expressions. This formulation not only offers a natural embodiment of the principles of belief revision as formulated by AGM [3], but also allows revisions based on imprecise observations. In addition, it enables features such as absorption of new conditional sentences and verification of counterfactual sentences and nested conditionals.

The lack of a mechanism for distinguishing causal relationships from other kinds of associations has been a serious deficiency in most nonmonotonic sys-

tems [96]. This problem is solved by augmenting the basic framework of ranking systems with a simple mechanism, called *stratification*, for the representation of causal relationships, actions, and changes. A new norm of consistency for knowledge bases containing causal rules was introduced, and applications to tasks of prediction and explanation were shown. The addition of stratification provides the necessary machinery for embodying belief updates and belief revision within the same framework.

6.2 Future Work

The next three sections sketch new directions for research into extending the semantical and computational framework presented in this dissertation.

6.2.1 Semantical Extensions

Section 4.6 briefly sketched how to interpret iterated (and embedded) conditionals using the ranking based semantics proposed in this dissertation. More work is required in order to characterize, in a manner similar to the postulates in [3, 65] for belief revision and update, the process of revising a knowledge base Δ with a conditional rule $\varphi \to \psi$. Of special interest are the cases where $\Delta \cup \{\varphi \to \psi\}$ is inconsistent. The final goal is the development of a meta-logic in which the connective $\to$ can be treated as a another connective in the underlying language. First steps can be found in [51], where system-Z is augmented to accept expressions of the form $\neg(\varphi \to \psi)$; "it is not the case that typically if φ then ψ". Semantically, $\neg(\varphi \to \psi)$ is interpreted as establishing that in the context of φ, the occurrence of ψ is as surprising or even more unlikely than the occurrence of $\neg\psi$. In terms of rankings $\neg(\varphi \to \psi)$ translates into:

$$\kappa(\varphi \land \psi) \geq \kappa(\varphi \land \neg\psi) \tag{6.1}$$

Procedures for testing consistency and answering queries requiring a polynomial number of satisfiability tests are also presented in [51].

Finally, the formulation in this dissertation is strictly propositional. Of primary interest are extensions to the first order case along the lines presented in [73].

6.2.2 Qualitative and Quantitative Information

The connection established in this dissertation between probability theory and qualitative forms of common sense inference provides a solid basis for combining qualitative information in the form of linguistic quantifiers, such as "likely", "very likely", "extremely likely", etc, with numerical probabilistic and statistical knowledge. The advantage of this proposal is that both qualitative and quantitative information can be processed coherently under a uniform semantical interpretation, in full conformity with the norms of probability calculus. Efforts should concentrate on developing an architecture where computation is performed in a parallel and distributed fashion and where precision is a function of the required urgency of the response (anytime response). The distributed algorithms developed by Pearl for Bayesian Networks [97] and the work by Hunter [61] on parallel belief revision provide a good starting point, applicable to cases where a unique stratified ranking can be established.[1]

This architecture will have an immediate impact in diagnosis systems and in the interpretation of sense data where the contributions of both qualitative and quantitative data are required, and anytime response is essential.

Autonomous planning systems are also likely candidates to benefit from such architecture. However, these systems not only require the ability to reason with defaults, evidence, and actions, but they also require the ability to reason about what is *desirable* and/or *difficult*, according to the consequences and costs of these actions. The trade-offs between actions, chances, and pay-offs have been studied thoroughly in decision theory [124, 64], and their application to AI has been emphasized recently [26, 125, 59]. In almost all formalisms proposed judgements about the likelihood of events is quantified by numerical probabilities and judgements about the desirability of action consequences are quantified by utilities, thus they are subject to the same criticisms (of the numerical approach) that motivated the work in this thesis.[2] An extension of the ranking formalism to include a qualitative abstraction of utilities should bring the computational and representational benefits that the approach in this dissertation presents for reasoning with default information.

---

[1] Hunter [61] adapts the algorithms in [97] for computing with Spohn's OCF.

[2] See [106] for an approach to default reasoning based on utilities, and [126] for a development of better representation languages.

6.2.3 Learning

By virtue of its probabilistic semantical basis, the framework proposed in this dissertation establishes a connection to learning, and enables us to ask not merely how to reason with defaults but also where default rules come from. It lays the theoretical foundations for learning systems that coherently extract conditional rules from raw observations, integrate them with rules transmitted linguistically, and further refine them to adapt to new changes in the environment.[3]

In Bayesian belief networks, the learning task separates nicely into two subtasks: Learning the parameters of the network (i.e., the conditional probabilities) for a given network topology and identifying the topology itself. These subtasks are clearly not independent because the set of parameters needed depends largely on the topology assumed, and conversely, the structure of the network is formally dictated by the joint distribution. Yet it is more convenient to execute the learning process in two separate phases: *structure learning* and *parameter learning*.[4] The topic of parameter learning is fairly well covered in the literature on estimation techniques [97, 119]. The task of structure learning is a more challenging one, and has received recent attention in [97, 39, 103], where methods and algorithms usually introduce assumptions of causality, and a preference towards simple structures. Given the relation between stratification and Bayes networks, it is to be expected that these methods can be adapted to the frameworks of Chapter 4 and 5. The parameters in a Bayes network correspond to the strength δ of the rules, and the topology of the network corresponds to the underlying graph structure Γ. The *do* operator introduced in Section 5.4 can be then used to uncover causal relations through the selective and controlled manipulation of events.

---

[3]In this section we regard *learning* as the task of finding a generic model of empirical data. In other words, learning can be thought of as the process of acquiring an effective internal representation for the persistent constraints in the world, i.e., generic facts and rules.

[4]The advantages are discussed in [97, Chapter 8].

APPENDIX A

Proofs

Since some of the proofs below refer to *unconfirmable* sets, we recall their definition:

Definition 2.16 A set $\Delta = D \cup S$ is said to be *unconfirmable* if one of the following conditions is true:

1. If D is nonempty, then there cannot be a defeasible sentence in D that is tolerated by Δ.

2. If D is empty (i.e., $\Delta = S$) then there must be a strict sentence in S which is non tolerated by Δ.

Essentially, unconfirmable sets are those that violate the conditions of Theorem 2.4 below.

Theorem 2.4 *Let* $\Delta = D \cup S$ *be a non-empty set of defeasible and strict sentences.* Δ *is p-consistent iff every non-empty subset* $\Delta' = D' \cup S'$ *of* Δ *complies with one of the following:*

1. *If* D' *is not empty, then there must be at least one defeasible sentence in* D' *tolerated by* Δ'.

2. *If* D' *is empty (i.e.,* $\Delta' = S'$*), each strict sentence in* S' *must be tolerated by* S'.

Proof of the *only if* **part:** We want to show that if there exists a non-empty subset of Δ which is unconfirmable, then Δ is not p-consistent. The proof is facilitated by introducing the notion of *quasi-conjunction* (see [2]): Given a set of defeasible sentences $D = \{\phi_1 \rightarrow \psi_1, \ldots, \phi_n \rightarrow \psi_n\}$ the *quasi-conjunction* of D is the defeasible sentence,

$$C(D) = [\phi_1 \vee \ldots \vee \phi_n] \rightarrow [(\phi_1 \supset \psi_1) \wedge \ldots \wedge (\phi_n \supset \psi_n)] \tag{A.1}$$

The quasi-conjunction $C(D)$ bears interesting relations to the set D. In particular, if there is a defeasible sentence in D which is tolerated (by D) by some

119

model ω, $C(D)$ will be verified by ω. This is so because the verification of at least one sentence of D by ω guarantees that the antecedent of $C(D)$ (i.e. the formula $[\phi_1 \vee \ldots \vee \phi_n]$ in Eq. (A.1)) is satisfied by ω, and the fact that no sentence in D is falsified guarantees that the consequent of $C(D)$ (i.e. the formula $[(\phi_1 \supset \psi_1) \wedge \ldots \wedge (\phi_n \supset \psi_n)]$ in Eq. (A.1)) is also satisfied by ω. Similarly, if at least one sentence of D is falsified by a model ω', its quasi-conjunction is also falsified by ω' since in this case, the consequent of $C(D)$ is not satisfied by ω' (at least one of the material implication in the conjunction is falsified by ω'). Additionally, let $U_p(C(D)) = 1 - P(C(D))$ (the *uncertainty* of $C(D)$) where $P(C(D))$ is the probability assigned to the quasi-conjunction of D according to Eq. (2.4), then, it is shown in [1] that the uncertainty of the quasi-conjunction of D is less or equal to the sum of the uncertainties of each of the sentences in D, i.e. $U_p(C(D)) \leq \sum_i (1 - P(\psi_i|\phi_i))$ where the sum is taken over all $\phi_i \rightarrow \psi_i$ in D.

We are now ready to proceed with the proof. Let $\Delta' = D' \cup S'$ be a nonempty subset of Δ where D' is a subset of D and S' is a subset of S. If Δ' is unconfirmable then one of the following cases must occur:

Case 1.- S' is empty and D' is unconfirmable[1]. In this case, the quasi-conjunction for D' is not verifiable; from Eq. (2.4), we have that for any P which is proper for $C(D')$, $P(C(D')) = 0$ and $U_p(C(D')) = 1$. It follows, by the properties of the quasi-conjunction outlined above that $\sum_i (1 - P(\psi_i'|\phi_i'))$ over all $\phi_i' \rightarrow \psi_i'$ in D' is at least 1. If the number of sentences in D' is $n \geq 1$, then,

$$n - \sum_{i=1}^{n} P(\psi_i'|\phi_i') \geq 1 \qquad (A.2)$$

$$\sum_{i=1}^{n} P(\psi_i'|\phi_i') \leq n - 1 \qquad (A.3)$$

which implies that at least one sentence in D' has probability smaller than $1 - \frac{1}{n}$. Hence, it is impossible to have $P(\psi_i'|\phi_i') \geq 1 - \varepsilon$, for every $\varepsilon > 0$, for every defeasible sentence $\phi_i' \rightarrow \psi_i' \in D'$. Thus, Δ is p-inconsistent.

Case 2.- D' is empty. If S' is unconfirmable, then there must be at least one sentence $\phi' \Rightarrow \sigma' \in S'$ such that no model ω' verifies $\phi' \Rightarrow \sigma'$ without falsifying another sentence in S'. We show by contradiction that there is no probability assingment P to the sentences in S' such that $P(\sigma|\phi) = 1$ for all $\phi \Rightarrow \sigma \in S'$ and P is proper for every sentence in S'. Assume there exists such a P. From Eq. (2.4)

$$P(\sigma|\phi) = \frac{\sum_{\omega \models \varphi \wedge \sigma} P(\omega)}{\sum_{\omega \models \varphi \wedge \sigma} P(\omega) + \sum_{\omega \models \varphi \wedge \neg\sigma} P(\omega)} = 1 \qquad (A.4)$$

---

[1]This case is covered by Theorem 1.1 in [2].

120

which immediately implies that if a model ω'' falsifies any sentence $\varphi'' \Rightarrow \sigma'' \in S'$ (including $\phi' \Rightarrow \sigma'$), then $P(\omega'')$ must be zero, else $P(\sigma''|\varphi'')$ will not equal 1. Thus, $P(\omega') = 0$ for every ω' verifying $\phi' \Rightarrow \sigma'$ since ω' must falsify another sentence in S'. But then either $P(\sigma'|\varphi') = 0$, or P is not proper for $\phi' \Rightarrow \sigma'$: A contradiction. We conclude that if S' is unconfirmable then Δ is p-inconsistent.

Case 3.- Neither D' nor S' are empty and Δ' is unconfirmable. That is, either the quasi-conjunction $C(D')$ is not verifiable or every ω' that verifies a defeasible sentence in D' falsifies at least one sentence in S'. The first situation will lead us back to case 1 while the second to a contradiction similar to case 2 above. In either case, Δ is not p-consistent.

Proof of the *if* **part:** Assume that every non-empty subset of $\Delta = D \cup S$ complies with the conditions of Theorem 2.4. Then the following two constructions are feasible:

- We can construct a finite "nested decreasing sequence" of non-empty subsets of Δ, namely $\Delta_1, \ldots, \Delta_m$, $(\Delta = \Delta_1)$, and an associated sequence of truth assignments $\omega_1, \ldots, \omega_m$ such that ω_i satisfies all the sentences in Δ_i and verifies at least one one defeasible sentence in Δ_i, and the sets in the sequence present the following characteristics:

  1. Δ_{i+1} is the proper subset of Δ_i consisting of all the sentences of D_i not verified by ω_i, for $i = 1, \ldots, m - 1$, plus the sentences in S.

  2. All sentences in D_m are verified by ω_m.

- We can construct a sequence $\omega_{m+1}, \ldots, \omega_n$ that will *confirm* $\Delta_{m+1} = S$. That is, the sequence $\omega_{m+1}, \ldots, \omega_n$ will verify every sentence in S without falsifying any. We will associate with $\omega_{m+1}, \ldots, \omega_n$ the "nested decreasing sequence" $\Delta_{m+1}, \ldots, \Delta_n$ where Δ_{i+1} is the proper subset of Δ_i consisting of all the sentences of S_i not verified by ω_i for $i = m + 1, \ldots, n$.

We can now assign probabilities to the truth-assignments $\omega_1, \ldots, \omega_n$ in the following way:

For $i = 1, \ldots, n - 1$

$$P(\omega_i) = \varepsilon^{i-1}(1 - \varepsilon) \tag{A.5}$$

and

$$P(\omega_n) = \varepsilon^{n-1} \tag{A.6}$$

121

We must show that, in fact, every $\varphi \to \psi$ in D obtains $P(\psi|\varphi) \geq 1 - \varepsilon$ and that every $\phi \Rightarrow \sigma$ in S obtains $P(s) = 1$. Since every $\varphi \to \psi$ is verified in at least one of the member of the sequence $\Delta_1, \ldots, \Delta_n$, using Eq. (2.4) we have that for $i < n$:

$$P(\psi_i|\phi_i) \geq \frac{\varepsilon^{i-1}(1 - \varepsilon)}{\varepsilon^{i-1}(1 - \varepsilon) + \varepsilon^i(1 - \varepsilon) + \cdots + \varepsilon^{n-1}} = 1 - \varepsilon \qquad (A.7)$$

and $P(\psi_n|\phi_n) = 1$ if it is only verified by the last model when S is originally empty. Finally, since no $\phi \Rightarrow \sigma$ in S is ever falsified by the sequence of truth assignments $\omega_1, \ldots, \omega_n$ and each and every $\phi \Rightarrow \sigma$ is verified at least once, it follows from Eq (2.4) and the process by which we assigned probabilities to $\omega_1, \ldots, \omega_n$ that indeed $P(\sigma|\phi) = 1$ for every $\phi \Rightarrow \sigma \in S$. $\square$

Corollary 2.5 $\Delta = D \cup S$ *is p-consistent iff we can build an ordered partition of* $D = [D_1, D_2, \ldots, D_n]$ *where:*

  1. *For all* $1 \leq i \leq n$, *each sentence in* D_i *is tolerated by* $S \cup_{j=i+1}^{j=n} D_j$.

  2. *Every sentence in* S *is tolerated by* S.

Proof: If Δ is p-consistent, by Theorem 2.4 we must be able to find a tolerated defeasible sentence in every subset $\Delta' = D' \cup S'$ (of Δ) where D' is nonempty, and it follows that the construction of the ordered partition $D = [D_1, D_2, \ldots, D_n]$ is possible. Similarly, by Theorem 2.4, if Δ is p-consistent every strict sentence in S must be tolerated by S. On the other hand, if both conditions in the corollary hold, we use the set of models (ω_i) that renders the sentences in each D_i tolerated by the set $S \cup_{j=i+1}^{j=n} D_j$ to construct a high probability model for Δ, following the probability assignments of Eqs. A.5 and A.6. $\square$

Theorem 2.8 *If* Δ *is p-consistent,* Δ *p-entails* $\varphi' \to \psi'$ *iff* $\phi' \to \neg\psi'$ *is substantively inconsistent with respect to* Δ.

Proof of the *only if* **part:** (If Δ p-entails $\varphi' \to \psi'$ then $\phi' \to \neg\psi'$ is substantively inconsistent with respect to Δ.) Let $\Delta \models_p \varphi' \to \psi'$. From the definition of p-entailment (Def. 2.7), for all $\varepsilon > 0$ there exists a $\delta > 0$ such that for all $P \in \mathcal{P}_{\Delta, \delta}$ which are proper for Δ and $\varphi' \to \psi'$, $P(\neg\psi'|\phi') \leq \varepsilon$. This means that for all proper probability assignments P for Δ and $\varphi' \to \psi'$ [2], the sentence $\phi' \to \neg\psi'$ gets an arbitrarily low probability whenever all defeasible sentences in Δ can be assigned arbitrarily high probability and all strict sentences in Δ can be assigned probability equal to 1. Thus $\phi' \to \neg\psi'$ is substantively inconsistent with respect to Δ.

---

[2]Note that from the definition of p-entailment there must exists at least one P proper for Δ and $\varphi' \to \psi'$.

Proof of the *if* **part:** (If $\phi' \to \neg\psi'$ is substantively inconsistent with respect to Δ then Δ p-entails $\varphi' \to \psi'$.) Let $\phi' \to \neg\psi'$ be substantively inconsistent with respect to Δ. From Theorem 2.4, we know that there must be a subset Δ' of $\Delta \cup \{\phi' \to \neg\psi'\}$ that is unconfirmable. Furthermore, since Δ is p-consistent, $\Delta' = \Delta'' \cup \{\phi' \to \neg\psi'\}$. Let $\mathcal{P}_S$ stand for the set of probability distributions that are proper for Δ and $\phi' \to \neg\psi'$ such that if $P \in \mathcal{P}_S$, then $P(\sigma|\phi) = 1$ for all $\phi \Rightarrow \sigma$ in Δ [3]. We will consider two cases depending on the structure of Δ'':

Case 1.- Δ'' does not include any defeasible sentences. From Theorem 2.4, we know that $\phi' \to \neg\psi'$ cannot be tolerated by Δ'' for otherwise Δ' wouldn't be inconsistent. It follows from Eq. 2.4 (probability assignment) that $P(\neg\psi'|\phi') = 0$ for all $P \in \mathcal{P}_S$. Thus, $P(\psi'|\phi') = 1$ in all $P \in \mathcal{P}_S$ and since any probability distribution that is in $\mathcal{P}_{\Delta,\varepsilon}$ must also belong to $\mathcal{P}_S$, it follows from the definition of p-entailment that $\Delta \models_p \varphi' \to \psi'$.

Case 2.- Δ'' includes defeasible and a possible empty set of strict sentences. Since $\Delta'' \cup \{\phi' \to \neg\psi'\}$ is unconfirmable, we have from the proof of Theorem 1, that for all probability distributions $P \in \mathcal{P}_S$:

$$\sum_{\varphi \to \psi \in \Delta''} U_p(\varphi \to \psi) + U_p(\phi' \to \neg\psi') \geq 1 \tag{A.8}$$

which implies that

$$\sum_{\varphi \to \psi \in \Delta} U_p(\varphi \to \psi) \geq 1 - U_p(\phi' \to \neg\psi') = U_p(\varphi' \to \psi') \tag{A.9}$$

Since $U_p(\varphi \to \psi) = 1 - P(\varphi \to \psi)$ and $U_p(\varphi' \to \psi') = 1 - P(\varphi' \to \psi')$, Eq. (A.9) says that $1 - P(\varphi' \to \psi')$ can be made arbitrarily small by requiring the values $1 - P(\varphi \to \psi)$ for $\varphi \to \psi \in D$ to be sufficiently small and the values of $P(\sigma|\phi)$ to be 1 for all $\phi \Rightarrow \sigma \in S$. This is equivalent to say that $\Delta \models_p \varphi' \to \psi'$. $\square$

Theorem 2.10 *If* $\Delta = D \cup S$ *is p-consistent,* Δ *strictly p-entails* $\phi' \Rightarrow \sigma'$ *iff* $S \cup \{\varphi' \to True\}$ *is p-consistent and there exists a subset* S' *of* S *such that* $\varphi' \Rightarrow \neg\sigma$ *is not tolerated by* S'.

Proof It follows from the proof of Theorem 2.8 (see case 1 of the *if* part). $\square$

Lemma A.1 *TEST_CONSISTENCY constitutes a decision procedure for testing the p-consistency of a set* Δ *of conditional sentences.*

---

[3] We know that $\mathcal{P}_S$ is not empty since $\Delta \cup \{\phi' \to True\}$ must be p-consistent according to Def. 2.6. In the case where Δ does not contain any strict sentences, $\mathcal{P}_S$ simply denotes all probability distributions that are proper for $\Delta \cup \{\phi' \to True\}$.

Proof: If the procedure stops at either line 4 or line 9 an unconfirmable subset is found, and by Theorem 2.4 the set of sentences is p-inconsistent. If on the other hand, the procedure reaches line 10, the order in which the sentences are tolerated can be used to build a high probability model for Δ using the construction (of the "nested decreasing sequence") in the proof of Theorem 2.4, and Δ must therefore be p-consistent. $\square$

Theorem 2.13 *The worst case complexity of testing consistency (or entailment) is bounded by $[\mathcal{PS} \times (\frac{|D|^2}{2} + |S|)]$ where $|D|$ and $|S|$ are the number of defeasible and strict sentences respectively, and $\mathcal{PS}$ is the complexity of propositional satisfiability for the material counterpart of the sentences in the database.*

Proof: Given that TEST_CONSISTENCY constitutes a decision procedure for p-consistency (see Lemma A.1 above), a complexity bound for this procedure will be an upper bound for the problem of deciding p-consistency. To assess the time complexity of TEST_CONSISTENCY, note that the WHILE-loop of line 6 will be executed $|S|$ times in the worst case, and each time we must do at most $\mathcal{PS}$ work to test the satisfiability of $S - s$; thus, its complexity is $|S| \times \mathcal{PS}$. In order to *find* a tolerated sentence $d : \phi \rightarrow \psi$ in D', we must test at most $|D'|$ times (once for each sentence $d \in D'$) for the satisfiability of the conjunction of $\phi \wedge \psi$ and the material counterparts of the sentences in $S \cup D' - \{d\}$. However, the size of D' is decremented by at least one sentence in each iteration of the WHILE-loop in line (2), therefore the number of times that we test for satisfiability is $|D| + |D| - 1 + |D| - 2 + \ldots + 1$ which is bounded by $\frac{|D|^2}{2}$. Thus, the overall time complexity is $O[\mathcal{PS} \times (\frac{|D|^2}{2} + |S|)]$. $\square$

Theorem 2.24 *If the set Δ is acyclic and of* Horn *form, wp_2-implication can be decided in polynomial time.*

Proof: The proof of this theorem requires a short review of some results from [25], since the procedure for deciding wp_2-implication is based on one of the algorithms presented in that paper. Given a set $\mathcal{H}$ of Horn clauses, Dowling and Gallier define an auxiliary graph $G_{\mathcal{H}}$ to represent the set $\mathcal{H}$, and reduce the problem of finding a truth assignment satisfying the sentences in $\mathcal{H}$, to that of finding a *pebbling* on the graph using a breadth first strategy. We first describe these concepts more precisely and then apply them to the problem at hand:

Definition A.2 ([25]) Given a set $\mathcal{H}$ of Horn clauses, $G_{\mathcal{H}}$ is labeled directed graph with $N + 2$ nodes (a node for each propositional letter occurring in $\mathcal{H}$, a node for **true** and a node for **false**) and a set of labels $[M]$. It is constructed with i taking values in $[M]$ as follows depending of the form of the ith Horn formula in $\mathcal{H}$:

1. If it is a positive literal q, there is an edge from **true** to q labeled i.

2. If it is of the form $\neg p_1 \vee \ldots \vee \neg p_n$, there are n edges from $p_1, \ldots, p_n$ to **false** labeled i.

3. If it is of the form $\neg p_1 \vee \ldots \vee \neg p_n \vee q$, there are n edges from $p_1, \ldots, p_n$ to q labeled i.

A node q in $G_{\mathcal{H}}$ can be *pebbled* if and only if for some label i, all sources of incoming edges labeled i are pebbled. The node **true** is considered to be pebbled. A pebbled path is a path on the graph such that all its nodes are pebbled. Given the correspondence between a Horn rule h_i and the set of i-labeled edges in the graph we are going to use both terms (edge and rule) indistinctively. Thus, eliminating a rule h_i should be understood as removing the set of i-labeled edges from the graph. Similarly a pebbled rule will indicate that the associated nodes in the graph are pebbled etc. A graph $G_{\mathcal{H}}$ is considered to be *completely* pebbled, if and only if all nodes that remain unpebbled have at least one incoming edge with a source that cannot be pebbled; i.e., there cannot be a pebbled path from **true** to that node.

Lemma A.3 ([25]) *Let $\mathcal{H}$ be a set of Horn clauses and let $G_{\mathcal{H}}$ be its associated graph, $\mathcal{H}$ is unsatisfiable iff there is a pebbling in $G_{\mathcal{H}}$ from* true *to* false

This lemma and the existence of an $O(N^2)$ algorithm for deciding satisfiability are proven in [25] (N represents the number of occurrences of literals in the set of clauses). We now prove a couple of lemmas regarding a polynomial procedure for deciding whether a conditional sentence x is weakly inconsistent with respect to a set Δ. Recall that by the definition of wp$_2$-implication (Def. 2.22) once we have identified a sentence as weakly inconsistent, its negation is wp$_2$-implied. The lemma below shows a simple test for deciding whether a particular horn sentence h is essential for the unsatisfiability of some set $\mathcal{H}$:

Lemma A.4 *Let $G_{\mathcal{H}}$ be an acyclic graph representing the set $\mathcal{H}$ of Horn clauses. Assume that $\mathcal{H}$ is unsatisfiable and that $G_{\mathcal{H}}$ is completely pebbled. Let $h \in \mathcal{H}$ be a Horn clause such that both the antecedent and consequent of h are pebbled in $G_{\mathcal{H}}$, and assume that there is a pebbled path from the consequent of h to* false*. Then there exists a nonempty subgraph $G'_{\mathcal{H}}$ of $G_{\mathcal{H}}$ containing h such that $G'_{\mathcal{H}}$ is unsatisfiable but $G'_{\mathcal{H}} - \{h\}$ is satisfiable.*

We show the correctness of this lemma by constructing the graph $G'_{\mathcal{H}}$. The idea is to eliminate from $G_{\mathcal{H}}$ all the alternative pebbled paths to **false**, and leave $G'_{\mathcal{H}}$

with only the path that goes through the rule h, together with those necessary to render this path pebbled. First, we select one pebbled path from **true** to **false** that goes through h (by the assumptions of the lemma, we know that there is at least one.) Next, we eliminate any rule that reaches **false** directly (i.e. of form 2 in Def. A.2) that is not in the selected path. We now traverse the selected path "backwards" from **false** to the node representing the consequent of h, and remove any incoming edges are not necessary to render this path pebbled. Note that we can guarantee to have eliminated alternative paths to **false**. The only possibility for this construction to fail is if we would have removed some paths that pebble the antecedents of h (in which case $G'_{\mathcal{H}}$ would be satisfiable), but this can only happen if there is a cycle in the graph involving h, and this possibility is ruled out by the assumptions of acyclicity. Since to complete the pebbling of a graph is no worse than testing for satisfiability, and searching for a pebbled path from one node to another can also be done by a breadth first search algorithm it follows that the test outlined in Lemma A.4 can be performed in polynomial time. This test constitutes the basis of a procedure for deciding weakly inconsistency:

Lemma A.5 *Given a set Δ which of Horn form and acyclic, to decide whether a sentence is weakly inconsistent with respect to Δ requires polynomial time.*

Given a set Δ and a sentence x, we first apply the consistency test of Section 2.4 to $\Delta \cup \{x\}$ in order to find an unconfirmable subset Δ_u. If none can be found or the sentence x does not belong to Δ_u, we can assert that x is *not* weakly inconsistent with respect to Δ. In the first case Δ is consistent, and in the second case x does not belong to any inconsistent subset of $\Delta \cup \{x\}$. Once Δ_u is found (and $x \in \Delta_u$), we systematically complete the pebbling of the associated graph G_{Δ_u} starting from each one of the antecedents of the sentences in Δ_u. If in one of these pebblings, the sentence x complies with the requirements of the test outlined in Lemma A.4, then x is weakly inconsistent. Note that all the steps involved require polynomial time with respect to N (i.e. the number of occurrences of literals in the set of clauses), and since once we have a procedure for deciding whether a sentence is weakly inconsistent we have a procedure for wp$_2$-implication (see Def. 2.22), we have essentially proven Theorem 2.24. $\square$

We remark that these results are not relevant only to nonmonotonic reasoning but to any application involving propositional entailment.

Theorem 3.3 *A set D is* consistent *(in the sense of Def. 3.2) iff D is* p-consistent.

Proof: By Theorem 2.4, if D is p-consistent then by there exist at least one tolerated rule in every nonempty subset $D' \subseteq D$. It follows that we can use the

same construction used in the proof of Theorem 2.4 (see Eqs. A.5 and A.6), and build a probability function parameterized on ε such that for each $\varphi_i \to \psi_i \in D$,

$$\lim_{\varepsilon \to 0} P_\varepsilon(\psi_i | \varphi_i) = 1 \qquad (A.10)$$

(see Eq. A.7), and it follows that D is consistent according to Definition 3.2.

On the other hand, by Theorem 2.4, if D is not p-consistent, then there exist a subset D' where no default rule is tolerated. We show that if this is the case, there cannot be an admissible ranking for D. Following Proposition 3.8 D is inconsistent (according to Definition 3.2), and the other direction of Theorem 3.3 holds. We reason by contradiction: Assume that there is no tolerated rule in $D' \subseteq D$ and there is an admissible ranking κ' for D. Let us define a characteristic possible world for a rule to be a possible world with minimal ranking verifying the rule. Since there is no tolerated rule in D', we know that any characteristic possible world ω_1 for rule $r_1 \in D'$ must falsify another rule $r_2 \in D'$. By the admissibility of κ' the following must hold

$$\kappa'(\omega_2) < \kappa'(\omega_1) \qquad (A.11)$$

where ω_2 is a characteristic possible world for r_2. By the same token, ω_2 must falsify another rule in D', say r_3, and we can insert $\kappa'(\omega_3)$ [4] in the chain of Eq. A.11:

$$\kappa'(\omega_3) < \kappa'(\omega_2) < \kappa'(\omega_1) \qquad (A.12)$$

We can continue to expand the chain in this fashion and get,

$$\kappa'(\omega_n) < \kappa'(\omega_{n-1}) < \ldots < \kappa'(\omega_2) < \kappa'(\omega_1) \qquad (A.13)$$

Note that if at any point in the construction of this chain, a possible world falsifies a rule that has a characteristic possible world in the chain, we arrive at a contradiction since by the admissibility of κ', $\kappa'(\omega') < \kappa'(\omega'')$ but since both ω' and ω'' are characteristic possible worlds of the same rule it must be that $\kappa'(\omega') = \kappa'(\omega'')$. Moreover, given that D' is finite we are bound to encounter such contradiction. $\square$

Theorem 3.6 *A PPD consequence relation satisfies the* Logic, Cumulativity, Cases and Rational Monotony *rules of inference.*

Proof: Note that if $\varphi \vdash \psi$ then $P(\psi | \varphi) = 1$. Thus, each PPD consequence relation satisfies the *Logic* rule. From elementary probability equivalences,

$$P(\gamma | \varphi) = P(\gamma | \psi \wedge \varphi)P(\psi | \varphi) + P(\gamma | \neg\psi \wedge \varphi)P(\neg\psi | \varphi); \qquad (A.14)$$

---

[4]ω_3 is a characteristic possible world for r_3.

thus, $\lim_{\varepsilon \to 0} P_\varepsilon(\gamma|\varphi)$ approaches $\lim_{\varepsilon \to 0} P_\varepsilon(\gamma|\psi \wedge \varphi)$ as $\lim_{\varepsilon \to 0} P_\varepsilon(\psi|\varphi)$ approaches 1. Hence, each such relation satisfies *Cumulativity*. Again, from elementary probability equivalences we have $P(\gamma|\varphi \vee \psi) = P(\gamma|\varphi) + P(\gamma|\psi) - P(\gamma|\varphi \wedge \psi)$. Thus, $\lim_{\varepsilon \to 0} P_\varepsilon(\gamma|\varphi \vee \psi) \geq \lim_{\varepsilon \to 0} P_\varepsilon(\gamma|\varphi) + \lim_{\varepsilon \to 0} P_\varepsilon(\gamma|\psi) - 1$, and it follows that a PPD consequence relation also satisfies *Cases*. Finally, since $P(\neg\psi|\varphi) = P(\neg\psi|\gamma \wedge \varphi)P(\gamma|\varphi) + P(\neg\psi|\neg\gamma \wedge \varphi)P(\neg\gamma|\varphi)$, if $\lim_{\varepsilon \to 0} P_\varepsilon(\neg\psi|\varphi) = 0$ (i.e. $\varphi \mathrel{\vute} \psi$) and $\lim_{\varepsilon \to 0} P_\varepsilon(\gamma|\varphi) \neq 0$, it must be the case that $\lim_{\varepsilon \to 0} P_\varepsilon(\neg\psi|\gamma \wedge \varphi) = 0$ (i.e. $\gamma \wedge \varphi \mathrel{\vute} \psi$) and it follows that a PPD consequence relation also satisfies *Rational Monotony*. $\square$

Theorem 3.6 *Every PPD consequence relation can be represented as a ranked preferential model, and every ranked preferential model with a finite non-empty state space can be represented as a PPD consequence relation.*

Proof: We have shown (Theorem 3.5) that each PPD consequence relation satisfies *Logic, Cumulativity, Cases and Rational Monotony*, and hence by the representation theorem in [74] it can be represented as a ranked preferential model.

For the converse part, we employ essentially the same construction as used in Lemma 31 of [74], except we take pains to ensure the probability functions are polynomial in ε. Suppose we are given a ranked preferential model with n ranks, denoted by $R_1, \ldots, R_n$. Let a_i be the number of states in R_i, for $1 \leq i \leq n$. For each state s,[5] define

$$P_\varepsilon(s) = \begin{cases} [1 - \varepsilon - \ldots - \varepsilon^{n-1}]/a_1 & \text{for } s \text{ in } R_1 \\ \varepsilon^{i-1}/a_i & \text{for } s \text{ in } R_i, \ 2 \leq i \leq n \end{cases}$$

It is easy to see that this probability measure on states will yield a PPD with the same consequence relation as the given ranked preferential model. $\square$

Theorem 3.10 *Given a consistent D, $\phi \mathrel{\vute_p} \sigma$ iff $D \models_p \phi \to \sigma$.*

Proof: We recall the definition of p-entailment (Def. 2.7). Given a positive real number ε, we say that a probability measure P ε-satisfies a default rule $\varphi \to \psi$, if $P(\psi|\varphi) \geq 1 - \varepsilon$. According to Definition 2.7, a default $\phi \to \sigma$ is p-entailed by a set D if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that every probability measure that δ-satisfies each rule in D will ε-satisfy $\phi \to \sigma$. Now suppose that $\phi \to \sigma$ is p-entailed by a set D. Let $\varepsilon > 0$ be arbitrary, and let δ be such that

---

[5] For technical reasons, Lehmann and Magidor [74] define ranked preferential models in terms of a set of states S, and a function l mapping states $s \in S$ to possible worlds $\omega \in \Omega$. A state s will *satisfy* a formula φ if and only if $l(s) \models \varphi$. For the purposes of this proof we define P_ε as a probability measure on a set of states, and define $P_\varepsilon(\omega)$ as $P_\varepsilon(\omega) = \sum_{l(s)=\omega} P_\varepsilon(s)$. The rest of our definitions and results hold without further modifications.

if P δ-satisfies each default in D, then it ε-satisfies $\phi \rightarrow \sigma$. Let P_γ be a PPD admissible with D. Then for each default $\varphi_i \rightarrow \psi_i$ in D, P_γ δ-satisfies $\varphi_i \rightarrow \psi_i$ for sufficiently small γ. Since D is finite, we can find a single constant K, such that P_γ δ-satisfies every member of D for $\gamma < K$. Thus, P_γ ε-satisfies $\phi \rightarrow \sigma$ for sufficiently small γ. Since ε is arbitrary, we conclude that σ is probabilistically entailed by D given ϕ.

For the converse, suppose that $\phi \rightarrow \sigma$ is not p-entailed by D. From Theorem 2.8, either $D \cup \{\phi \rightarrow \neg\sigma\}$ is consistent, or $D \cup \{\phi \rightarrow True\}$ is inconsistent. Suppose first that $D \cup \{\phi \rightarrow True\}$ is inconsistent. Since D is consistent, the construction in the proof of Theorem 2.4 can be used to obtain a PPD admissible with D. Since $D \cup \{\phi \rightarrow True\}$ is inconsistent, this PPD does not satisfy $\phi \rightarrow True$, and hence $\phi \hspace{1pt}\vdash\hspace{-6pt}\sim \sigma$ cannot hold in its proper consequence relation. Assume now that $D \cup \{\phi \rightarrow \neg\sigma\}$ is consistent. Once again using the construction in the proof of Theorem 2.4, we get a PPD admissible with respect to $D \cup \{\phi \rightarrow \neg\sigma\}$. Clearly, the induced consequence relation cannot satisfy $\phi \rightarrow \sigma$. Thus, σ is not probabilistically entailed by D given ϕ. $\square$

Proposition 3.12 *If D is an MC-set then, for all defaults $r : \varphi \rightarrow \psi \in D$, $\varphi \hspace{1pt}\not\vdash\hspace{-6pt}\sim \psi$ is not in the consequence relation induced by $D - \{\varphi \rightarrow \psi\}$.*

Proof: Note that since D is an MC set then $D - \{\varphi_i \rightarrow \psi_i\}$ must also be an MC set. By the MC set property (Def. 3.11) for each rule $r_i : \varphi_i \rightarrow \psi_i \in D$, there exists a possible world ω_i such that ω_i falsifies only r_i and no other rule in D. Thus, $\kappa^*(\omega_i)$ for the set $D - \{\varphi_i \rightarrow \psi_i\}$ must be equal to zero (no rule in $D - \{\varphi_i \rightarrow \psi_i\}$ is violated by ω_i) and thus any possible world $\omega' \models \varphi_i \wedge \psi_i$, must comply with $\kappa^*(\omega') \geq \kappa^*(\omega_i)$. It follows that $\varphi_i \hspace{1pt}\not\vdash\hspace{-6pt}\sim \psi_i$ is not in the consequence relation induced by $D - \{\varphi_i \rightarrow \psi_i\}$. $\square$

In order to show Theorem 3.13, we require the following lemma:

Lemma A.6 *Given an MC set D, there exists at least one Z function that satisfies eq. 3.19.*

Proof: We define an operator $\mathcal{O}$ on Z functions by $\mathcal{O}(Z) = Z'$, where

$$Z'(r_i) \;=\; 1 + \min_{\Omega^+_{r_i}} [\sum_{r_j \in D^-_\omega} Z(r_j)] \quad 1 \leq i \leq n$$

A Z function satisfies eq. 3.19 iff it is a fixed point of $\mathcal{O}$.

Now define a sequence $\{Z_n\}$ of Z functions by

$$Z_1(r) = 1 \text{ for every rule } r$$

129

and

$$Z_{n+1} = \mathcal{O}(Z_n).$$

We will prove that, for every r, the sequence $\{Z_n(r)\}$ is both non-decreasing and bounded, and hence converges.

First we show that $Z_{n+1}(r) \geq Z_n(r)$ for every r. (We will abbreviate this as $Z_{n+1} \geq Z_n$.) To see this, note that if $Z \geq Z'$, then $\mathcal{O}(Z) \geq \mathcal{O}(Z')$. Clearly, $Z_2 \geq Z_1$. Applying $\mathcal{O}$ to both sides, we get $Z_3 \geq Z_2$. This process can be repeated, showing (by induction) that $Z_{n+1} \geq Z_n$, for every n.

Next, we use the partition of D introduced in Corollary 2.5 to show that $\{Z_n(r)\}$ is bounded above for every r. We can do this by induction on the tolerance set to which r belongs. Clearly it is true for rules in D_0, since among the verifiers for such rules are possible worlds with no violations. Assume it is true for r in D_i. Consider a rule r that belongs to D_{i+1}. There must be at least one verifier ω of r that violates only rules in D_i. According to the inductive hypothesis, therefore, $\sum_{r_j \in D_M^-} Z(r_j)$ will be bounded during every application of $\mathcal{O}$. It follows that $\{Z_n(r)\}$ is bounded.

Since $\{Z_n\}$ converges, we can define

$$Z = \lim_{n \to \infty} Z_n.$$

Clearly Z will be a fixed point of $\mathcal{O}$. $\square$

Relying on Lemma A.6 we can now define Z^* to be an arbitrarily chosen solution to Eq. 3.19. It will follow from Theorem 3.13 that this solution is unique.

Theorem 3.13 *Given an MC set D,* Procedure Z^*_order *computes the function Z^* defined by Eqs. 3.20 and 3.21.*

Proof: We first show that the relevant steps in Procedure Z^*_order are well defined. By the assumption that D is consistent, D_0 cannot be an empty set (steps 1 and 2): There must be at least one rule tolerated by D. By similar reasons, Ω cannot be empty in each iteration of the loop in step 3. By consistency we must be able to find a tolerated sentence in each nonempty subset of D. Finally, in the computation of Eq. 3.23, since ω only falsifies rules in $\mathcal{R}Z^+$, all Z for these rules are available.

We now show that $Z = Z^*$ for rules $r_0 \in D_0$. Since each r_0 is tolerated by D, there must be a possible world ω_0 (for each one of these rules), such that ω_0 verifies r_0 and ω_0 satisfies D. Thus, each one of these possible worlds does not falsify any rules in D, and $\kappa^*(\omega_0) = 0$. According to Eq. 3.21, $Z^*(r_0) = 1$ for those rules and that is precisely what is computed in step 2.

The proof proceeds by induction on the iterations of loop 3; we show that for every rule $r \in \mathcal{R}Z^+$, $Z(r) = Z^*(r)$ holds as an invariant. For the basis of the induction consider the first iteration: Since $\mathcal{R}Z^+ = D_0$, then for every $r_0 \in D_0$, $Z(r_0) = Z^*(r_0)$ holds as shown above. Our objective is to show that this equality holds for the rules inserted into $\mathcal{R}Z^+$ at step 3.(c). Note that since all the values $\kappa(\omega)$ for $\omega \in \Omega$ are computed from Z^*-values of rules in $\mathcal{R}Z^+$ (step 3.b Eq. 3.23), they must be equal to $\kappa^*(\omega)$. We define a characteristic possible world for a rule r to be the possible world ω_r with minimal ranking κ^* verifying r. Thus, $Z^*(r) = \min_{\omega \models \varphi \wedge \psi} \kappa^*(\omega) + 1 = \kappa^*(\omega_r) + 1$. We claim that $\kappa^*(\omega^*)$[6] is a characteristic possible world for the rules outside $\mathcal{R}Z^+$ it verifies. Suppose not: Assume that there is a possible world $\omega_{r'}$ such that $\omega_{r'}$ verifies some rule r' which ω^* also verifies, and $\kappa^*(\omega_{r'}) < \kappa^*(\omega^*)$. Note that $\omega_{r'}$ cannot belong to Ω since the value of $\kappa^*(\omega^*)$ is minimal with respect to the κ^* of possible worlds in Ω. It follows then that $\omega_{r'}$ must falsify a rule $r'' \notin \mathcal{R}Z^+$. Let $\omega_{r''}$ be a characteristic possible world for r'', then

$$\kappa^*(\omega_{r''}) < \kappa^*(\omega_{r'}) \tag{A.15}$$

Note that ω^* cannot verify r'', since otherwise

$$\kappa^*(\omega^*) < \kappa^*(\omega_{r'}) \tag{A.16}$$

a contradiction. By the same argument as above, $\omega_{r''} \notin \Omega$, and therefore it must falsify a rule $r''' \notin \mathcal{R}Z^+$. if $\omega_{r'''}$ is a characteristic possible world for r''' we have that

$$\kappa^*(\omega_{r'''}) < \kappa^*(\omega_{r''}) < \kappa^*(\omega_{r'}) \tag{A.17}$$

$\omega_{r'}$ cannot verify r'''; otherwise we get the contradiction

$$\kappa^*(\omega_{r'}) < \kappa^*(\omega_{r''}) < \kappa^*(\omega_{r'}) \tag{A.18}$$

and if ω^* verifies r''' we get the contradiction of Eq. A.16. $\omega_{r'''}$ cannot belong to Ω and therefore it must falsify another rule outside $\mathcal{R}Z^+$. However, given that D is finite, we cannot extend the "chain" of Eq. A.17 indefinitely, and therefore we are bound to get a contradiction in the form of Eq. A.16 or Eq. A.18. Since our only assumption was that $\kappa^*(\omega^*)$ is not a characteristic possible world for the rules it verifies, that assumption must be wrong. It follows then that the value of $Z(r)$ computed in step 3.c (Eq. 3.24) must be equal to Z^*. For the induction

---

[6]Recall that ω^* is a possible world in Ω with minimal value κ (see step 3.c in Procedure Z^*_order).

step assume that the invariant holds up till the nth iteration. Then by the same argument used in the basis of the induction, the $\kappa(\omega)$ for $\omega \in \Omega$ are equal to $\kappa^*(\omega)$, ω^* must be a characteristic possible world for the rules r outside of $\mathcal{R}Z^+$ that it verifies, and thus $Z(r) = \kappa^*(\omega^*) + 1 = Z^*(r)$. $\square$

Theorem 4.3 *A set Δ is consistent (in the sense of Def. 4.2) iff $\bar{\Delta}$ is p-consistent.*

Proof: By Theorem 2.4, Δ is p-consistent iff there exist at least one tolerated rule (by Δ') in every nonempty subset Δ' of Δ. We first show that if there exists a tolerated rule in every nonempty subset of Δ we can always produce an admissible ranking κ. Under the stated condition, we can construct the following ordered partition $(\Delta_0, \Delta_1, \ldots, \Delta_n)$ of Δ: Rules in Δ_0 are tolerated by Δ, rules in Δ_1 are tolerated by $\Delta - \Delta_0$ and so on (see Cor. 2.5). By Def. 2.3,[7] for each one of these Δ_j, there must exist a nonempty subset Ω_j of Ω (the set of all possible possible worlds), such that for each rule $r_j \in \Delta_j$ there must exists a possible world $\omega_j \in \Omega_j$, where ω_j verifies r_j and ω_j satisfies Δ if $j = 0$ and $\Delta - \{\Delta_0 \cup \ldots \cup \Delta_{j-1}\}$ otherwise. Thus, using these possible worlds (the possible worlds actually required to effectively build the partition of Δ), we define a partition $(\Omega_0, \Omega_1, \ldots, \Omega_n, \Omega_{n+1})$ of Ω, where each Ω_j contains possible worlds with the characteristics mentioned above, and Ω_{n+1} contains the possible worlds necessary to complete the partition. Let δ_i^* denote the highest δ among rules in set Δ_i. We now build, in a recursive fashion, an admissible ranking κ based on these two partitions in the following manner: If $\omega_0 \in \Omega_0$, set $\kappa(\omega_0) = 0$. Else if $\omega_j \in \Omega_j$, set $\kappa(\omega_j) = \kappa(\omega_{j-1}) + \delta_{j-1}^* + 1$. Note that each possible world $\omega_j \in \Omega_j$ is a characteristic possible world[8] of the rule $r_j \in \Delta_j$ it verifies, and the κ-minimal possible world falsifying any rule $r_j \in \Delta_j$ must belong to the set Ω_{j+1}. Thus, in order to guarantee the admissibility of κ, it is enough to show that for an arbitrary pair of possible worlds $\omega_j \in \Omega_j$ and $\omega_{j+1} \in \Omega_{j+1}$ the following relation holds:

$$\kappa(\omega_j) + \delta_j < \kappa(\omega_{j+1}) \tag{A.19}$$

where δ_j can be any δ among the rules in Δ_j. But this relation is guaranteed by the construction of κ since $\kappa(\omega_j) + \delta_j^* + 1 = \kappa(\omega_{j+1})$, where δ_j^* is the highest δ among the rules in Δ_j. Therefore κ is admissible.

To show the converse we reason by contradiction: Assume that there is no tolerated rule in $\Delta' \subseteq \Delta$ and there is an admissible ranking κ' for Δ (this part

---

[7]Rules with strength δ are verified, falsified, and therefore tolerated in the same way that rules without strength δ.

[8]Recall that a possible world ω^+ is said to be a *characteristic possible world* for rule $\varphi \to \psi$ relative to ranking κ, if $\kappa(\omega^+) = \min\{\kappa(\omega) : \omega \models \varphi \wedge \psi\}$.

of the proof is almost identical to the proof of Theorem 3.3, except for the δ's on the rules). Since there is no tolerated rule in Δ', we know that any characteristic possible world ω_1 for rule $r_1 \in \Delta'$ must falsify another rule $r_2 \in \Delta'$. By the admissibility of κ' the following must hold

$$\kappa'(\omega_2) + \delta_2 < \kappa'(\omega_1) \tag{A.20}$$

where ω_2 is a characteristic possible world for r_2. By the same token, ω_2 must falsify another rule in Δ', say r_3, and we can insert $\kappa'(\omega_3)$ [9] in the chain of Eq. A.20:

$$\kappa'(\omega_3) + \delta_3 < \kappa'(\omega_2) + \delta_2 < \kappa'(\omega_1) \tag{A.21}$$

We can continue to expand the chain in this fashion and get,

$$\kappa'(\omega_n) + \delta_n < \kappa'(\omega_{n-1}) + \delta_{n-1} < \ldots <$$
$$\kappa'(\omega_2) + \delta_2 < \kappa'(\omega_1) \tag{A.22}$$

Note that if at any point in the construction of this chain, a possible world falsifies a rule that has a characteristic possible world in the chain, we arrive at a contradiction since by the admissibility of κ', $\kappa'(\omega') + \delta' < \kappa'(\omega'')$ but since both ω' and ω'' are characteristic possible worlds of the same rule it must be that $\kappa'(\omega') = \kappa'(\omega'')$. Moreover, given that Δ' is finite we are bound to encounter such contradiction. $\square$

Proposition A.7 *The ranking function κ^+ is admissible.*

Proof: Given that $Z^+(r_i) = \min\{\kappa^+(\omega) : \omega \models \phi_i \wedge \psi_i\} + \delta_i$, we can re-write the conditions for admissibility (Eq. 4.6) as

$$Z^+(r_i) < \min\{\kappa^+(\omega) : \omega \models \varphi_i \wedge \neg\psi_i\} \tag{A.23}$$

Since $\kappa^+(\omega) = \max\{Z^+(r_i) : \omega \models \varphi_i \wedge \neg\psi_i\} + 1$, it follows that κ^+ is admissible. $\square$

Lemma A.8 *The ranking κ^+ is compact.*

Proof: By contradiction. Assume it is possible to lower $\kappa^+(\omega')$ of some possible world ω', where $\kappa^+(\omega') > 0$. From the definition of κ^+ (Def. 4.4, there must be a rule $r : \varphi \xrightarrow{\delta} \psi$ such that $\kappa^+(\omega') = Z^+(r) + 1$ (see Eq. 4.7), which implies that

$$\kappa^+(\omega') = \min\{\kappa^+(\omega) : \omega \models \varphi \wedge \psi\} + \delta + 1 \tag{A.24}$$

---

[9] ω_3 is a characteristic possible world for r_3.

133

Lowering the value of $\kappa^+(\omega')$ will violate Eq. A.24 which will imply the violation of Eq. 4.6 for rule r. $\square$

Theorem 4.7 *Every consistent* Δ *has a* unique compact ranking *given by* κ^+ '.

Proof: By Lemma A.8, κ^+ is compact. We show it is also unique. Suppose there exists some other compact ranking κ that differs from κ^+ in at least one possible world. We will show that if there exists an ω' such that $\kappa(\omega') < \kappa^+(\omega')$ then κ cannot be admissible, where if $\kappa(\omega') > \kappa^+(\omega')$, then κ cannot be compact. Assume $\kappa(\omega') < \kappa^+(\omega')$, let I be the lowest κ value for which such inequality holds, and let $\kappa^+(\omega') = J > I$. By the definition of κ^+ (Def. 4.4), we know that there is a rule $r : \varphi \xrightarrow{\delta} \psi$ such that Eq. A.24 holds, and as a consequence

$$\min\{\kappa^+(\omega) : \omega \models \varphi \wedge \psi\} = J - \delta - 1 \tag{A.25}$$

Since κ is assumed to be admissible, the following must hold for rule r

$$\kappa(\omega') \geq \min\{\kappa(\omega) : \omega \models \varphi \wedge \psi\} + \delta + 1 \tag{A.26}$$

Since $J > \kappa(\omega')$,

$$J > \min\{\kappa(\omega) : \omega \models \varphi \wedge \psi\} + \delta + 1 \tag{A.27}$$

If we subtract $\delta + 1$ from both sides of this inequality and use Eq. A.25 we get

$$\begin{aligned} \min\{\kappa^+(\omega) : \omega \models \varphi \wedge \psi\} > \\ \min\{\kappa(\omega) : \omega \models \varphi \wedge \psi\} \end{aligned} \tag{A.28}$$

But this cannot be since I was assumed to be the minimal value of κ for which this inequality can occur, and if $\min\{\kappa(\omega) : \omega \models \varphi \wedge \psi\} > I$, then κ is not admissible (see Eq. A.26).

Now assume that there is a non-empty set of possible worlds for which $\kappa(\omega) > \kappa^+(\omega)$, and let I be the lowest κ^+ value in which $\kappa(\omega') > \kappa^+(\omega')$ for some possible world ω'. We will show that κ cannot be compact, since it will be possible to reduce $\kappa(\omega')$ to $\kappa^+(\omega')$ while keeping constant the κ of all other possible worlds. From $\kappa^+(\omega') = I$ we know that ω' does not falsify any rule r with Z^+ rank higher than $I - 1$. Hence, we only need to watch whether the reduction of κ can violate rules r for which $Z^+(r) < I$. For every such rule there exists a possible world ω, such that ω verifies r and $\kappa^+(\omega) < I$. Since for all these possible worlds κ is assumed to be equal to κ^+ it follows that none of these possible worlds can be violated by reducing $\kappa(\omega')$ to $\kappa^+(\omega')$. $\square$

Theorem 4.9 *The function* Z *computed by* Z^+_order *complies with Definition 4.4, that is* $Z = Z^+$.

134

Proof: We first show that the relevant steps in Procedure Z^+_order are well defined. By the assumption that Δ is consistent, Δ_0 cannot be an empty set (steps 1 and 2): There must be at least one rule tolerated by Δ. By similar reasons, Δ^+ cannot be empty in each iteration of the loop in step 3. By consistency we must be able to find a tolerated sentence in each nonempty subset of Δ. Finally, in the computation of Eq. 4.10, since ω only falsifies rules in $\mathcal{R}Z^+$, all Z for these rules are available.

We now show that $Z = Z^+$ for rules $r_0 \in \Delta_0$. Since each r_0 is tolerated by Δ, there must be a possible world ω_0 (for each one of these rules), such that ω_0 verifies r_0 and ω_0 satisfies Δ. Thus, each one of these possible worlds does not falsify any rules in Δ, and $\kappa^+(\omega_0) = 0$. According to Eq. 4.8 in Def. 4.4, $Z^+(r_0) = \delta_0$ for those rules and that is precisely what is computed in step 2.

The proof proceeds by induction on the iterations of loop 3; we show that for every rule $r \in \mathcal{R}Z^+$, $Z(r) = Z^+(r)$ holds as an invariant. For the basis of the induction consider the first iteration: Since $\mathcal{R}Z^+ = D_0$, then for every $r_0 \in D_0$, $Z(r_0) = Z^+(r_0)$ holds as shown above. Our objective is to show that this equality holds for the rules r^* inserted into $\mathcal{R}Z^+$ at step 3.(c). Note that since all the values $\kappa^+(\omega_r)$ for ω_r in every Ω_r are computed from Z^+-values of rules in $\mathcal{R}Z^+$ (step 3.b Eqs. 4.10 and 4.11), they must be equal to $\kappa^+(\omega)$. As done in the proof of Theorem 3.13, let a characteristic possible world for a rule r be the possible world ω_r^* with minimal ranking κ^+ verifying r. Thus, $Z^+(r) = \min_{\omega \models \varphi \wedge \psi} \kappa^+(\omega) + \delta = \kappa^+(\omega_r^*) + \delta$. We claim that $\kappa^+(\omega_{r^*}^*)$[10] is a characteristic possible world for the rules outside $\mathcal{R}Z^+$ it verifies. Suppose not: Assume that there is a possible world ω_{r*} such that ω_{r*} verifies a rule r^* (that is inserted into $\mathcal{R}Z^+$ in step 3.c), and $\kappa^+(\omega_{r*}) < \kappa^+(\omega_{r^*}^*)$. Note that ω_{r*} must falsify a rule $r' \notin \mathcal{R}Z^+$. Otherwise the computation in Eq. 4.10 would not have used ω_{r*}^* but ω_r^* instead. Let $\omega_{r'}$ be a characteristic possible world for r', then

$$\kappa^+(\omega_{r'}) < \kappa^+(\omega_{r*}) \tag{A.29}$$

Note that ω_{r*}^* cannot verify r', since otherwise

$$\kappa^+(\omega_{r*}^*) < \kappa^+(\omega_{r*}) \tag{A.30}$$

a contradiction. If $\omega_{r'}$ does not verify the same rule r^* that ω_{r*}^* verifies, then $Z(r') \geq Z(r^*)$ by Step 3.c, and then by Eq. 4.11, $\kappa(\omega_{r*} > \kappa(\omega_{r*}^*$ which is a contradiction. Therefore $\omega_{r'}$ verifies the same r^*, and by the minimality of ω_{r*}^* among the worlds in Ω_{r*}, $\omega_{r'}$ must falsify a rule $r'' \notin \mathcal{R}Z^+$. If $\omega_{r''}$ is a characteristic

---

[10]Recall that r^* is a rule inserted into $\mathcal{R}Z^+$ in Step 3.c.

possible world for r'' we have that

$$\kappa^+(\omega_{r''}) < \kappa^+(\omega_{r'}) < \kappa^+(\omega_{r*}) \tag{A.31}$$

ω_{r*} cannot verify r''; otherwise we get the contradiction

$$\kappa^+(\omega_{r*}) < \kappa^+(\omega_{r'}) < \kappa^+(\omega_{r*}) \tag{A.32}$$

and if ω_{r*}^* verifies r'' we get the contradiction of Eq. A.30. By similar arguments as before $\omega_{r''}$ must falsify another rule outside $\mathcal{R}Z^+$. However, given that Δ is finite, we cannot extend the "chain" of Eq. A.31 indefinitely, and therefore we are bound to get a contradiction in the form of Eq. A.30 or Eq. A.32. Since our only assumption was that w_{r*}^* is not a characteristic possible world for the rules it verifies, that assumption must be wrong. It follows then that the value of $Z(r^*)$ computed in step 3.b (Eq. 4.10) must be equal to Z^+. For the induction step assume that the invariant holds up till the nth iteration. Then by the same argument used in the basis of the induction, the $\kappa(\omega_r)$ for $\omega_r \in \Omega_r$ are equal to $\kappa^+(\omega_r)$, the minimal $\kappa^+(\omega_{r*}^*)$ in Eq. 4.10 must be a characteristic possible world for the rules r^* outside of $\mathcal{R}Z^+$ that it verifies, and thus $Z(r^*) = \kappa^+(\omega_{r*}^*) + \delta_{r*} = Z^+(r^*)$. $\square$

Lemma 4.10 *Let $\Delta = \{r_i \mid r_i = \varphi_i \xrightarrow{\delta_i} \psi_i\}$ be a consistent set where the rules are sorted in nondecreasing order according to priorities $Z(r_i)$. Let $\kappa(M)$ be defined as in Eq. 4.7:*

$$\kappa(M) = \begin{cases} 0 & \text{if } M \text{ does not falsify any rule in } \Delta \\ \max_{M \models \varphi_i \wedge \neg \psi_i}[Z(r_i)] + 1 & \text{otherwise.} \end{cases} \tag{A.33}$$

Then, for any wff ϕ, $\kappa(\phi)$ can be computed in $O(\log |\Delta|)$ propositional satisfiability tests.

Proof: The idea is to perform a binary search on Δ to find the lowest $Z(r)$ such that there is a model for ϕ that does not violate any rule r' with priority $Z(r') \geq Z(r)$. We first divide Δ into two roughly equal sections: top–half (r_{mid} to r_{high}) and bottom–half (r_{low} to r_{mid}). Then we examine the top–half: If the wff $\alpha = \phi \wedge_{j=mid}^{j=n} \varphi_j \supset \psi_j$ is satisfiable, then there exists a model for ϕ that does not violate any rule in this top–half. It follows that $Z(r_{mid}) + 1$ is an upper bound on the value of $\kappa(\phi)$, and the binary search is continued iteratively in the bottom–half. If, on the other hand, α is not satisfiable, then the maximum $Z(r_i)$ for any model for ϕ must be in the top–half, and the search is continued there. Eventually, the set in which the search is conducted is reduced to one rule, and we can determine the value of $\kappa(\phi)$ with one more satisfiability test. $\square$

Lemma 4.11 *The value of $Z(\phi \xrightarrow{\delta} \sigma)$ in Eq. 4.10 can be computed in $O(\log|\mathcal{R}Z^+|)$ satisfiability tests.*

Proof: Let Δ' in Step 3(a) be equal to $\{\varphi_i \xrightarrow{\delta_i} \psi_i\}$, and let the wff α be equal to $\sigma \wedge \phi \wedge_i \varphi_i \supset \psi_i$ where i ranges over all the rules in Δ'. Note that since any world M_r in $\mathcal{M}_r$ is a model for $\sigma \wedge \phi$ and does not violate any rule in Δ', it follows that $M_r \in \mathcal{M}_r$ iff $M_r \models \alpha$. Then, since $\kappa(\alpha) = \min_{M_r \in \mathcal{M}_r} \kappa(M_r)$, $Z(\phi \xrightarrow{\delta} \sigma)$ must be equal to $\kappa(\alpha) + 1 + \delta$. Thus, once Δ' is sorted, by Lemma 4.10, $\kappa(\alpha)$ can be computed in $O(\log|\mathcal{R}Z^+|)$ satisfiability tests which proves Lemma 4.11. $\square$

Theorem 4.12 *Given a consistent Δ, the computation of the ranking Z^+ requires $O(|\Delta|^2 \times \log|\Delta|)$ satisfiability tests.*

Proof: Step 1 requires at most $|\Delta|$ satisfiability tests and is performed once, while Step 2 takes at most $|\Delta|$ data assignments. Step 3(a) again requires $O(|\Delta|)$ satisfiability tests. Computing Eq. 4.10 in Step 3(b) can be done in $O(\log|\mathcal{R}Z^+|)$ satisfiability tests according to Lemma 4.11,[11] and since it will be executed at most $O(|\Delta|)$ times, it requires a total of $O(|\Delta| \times \log|\Delta|)$ satisfiability tests. Step 3(c) is a minimum search which can be done in conjunction with the computation of Eq. 4.10, since we only need to keep the minimum of such values. It involves $|\Omega|$ data comparisons. Loop 3 is performed at most $|\Delta| - |\Delta_0|$ times, hence the whole computation of the priorities Z^+ on rules requires a total of $O(|\Delta|^2 \times \log|\Delta|)$ satisfiability tests. $\square$

Theorem 4.14 *Let $r1 : \varphi \xrightarrow{\delta_1} \psi$ and $r2 : \phi \xrightarrow{\delta_2} \sigma$ be two rules in a consistent Δ such that*

1. $\varphi \mathrel{\vbox{\hbox{$\scriptstyle\vdash$}\hbox{$\scriptstyle p$}}} \phi$ (i.e., φ is more specific than ϕ).

2. There is no model satisfying $\varphi \wedge \psi \wedge \phi \wedge \sigma$ (i.e., r_1 conflicts with r_2).

Then $Z^+(r1) > Z^+(r2)$ independently of the values of δ_1 and δ_2.

Proof: If $\varphi \mathrel{\vdash\!\!\!\sim} \phi$ is in every consequence relation of every κ admissible with Δ then (by Prp. 3.8) the following constraint must hold in all these κ-rankings (including κ^+):

$$\kappa(\varphi \wedge \phi) < \kappa(\varphi \wedge \neg\phi) \tag{A.34}$$

---

[11]Note that we need $\mathcal{R}Z^+$ to be sorted, nondecreasingly, with respect to the priorities Z. This requires that the initial values inserted to $\mathcal{R}Z^+$ in Step 2 of **Procedure** Z^+_order be sorted — $O(|\Delta_0|^2)$ data comparisons — and that the new Z-value in Step 3(c) be inserted in the right place — $O(|\mathcal{R}Z^+|)$ data comparisons. We are assuming that the cost of each of these operations is much less than that of a satisfiability test.

Thus, any characteristic possible world $\omega_{r_1}^+$ for r_1 must render ϕ (the antecedent for r_2) true, and since there is no possible world such that both rules are verified (condition 2 in the theorem above), all $\omega_{r_1}^+$ must also falsify r_2. From Def. 5 (Eqs. 4.7 and 4.8): $\kappa^+(\omega_{r_1}^+) \geq Z^+(r_2) + 1$, and $Z^+(r_1) = \kappa^+(\omega_{r_1}^+) + \delta_2$. It follows that $Z^+(r_1) > Z^+(r_2)$. Note that the characteristic possible world for r_2 cannot in turn falsify r_1 since this will preclude the existence of an admissible ranking κ and Δ was assumed to be consistent. $\square$

Theorem 5.2 *Given a network Δ, let $\mathcal{O}_1(\mathcal{X})$ and $\mathcal{O}_2(\mathcal{X})$ be two orderings of the elements in $\mathcal{X}$ according to Δ. If κ is stratified for Δ under $\mathcal{O}_1(\mathcal{X})$, then κ is stratified for Δ under $\mathcal{O}_2(\mathcal{X})$.*

Proof: Let $\mathcal{L} = \{X_1, \ldots, X_n\}$ be the set of literals variables in the language taking values from the atomic propositions $\mathcal{X} = \{x_1, \ldots, x_n\}$. Let $\mathcal{O}_1 = [Y_1, \ldots, Y_n]$ and $\mathcal{O}_2 = [Z_1, \ldots, Z_n]$, where $\{X_1, \ldots, X_n\} = \{Y_1, \ldots, Y_n\} = \{Z_1, \ldots, Z_n\}$ and $[]$ denotes sequences. We will show that for $1 \leq i \leq n$

$$\kappa(Z_i \wedge \ldots \wedge Z_1) = \sum_{j=1}^{j=i} \kappa(Z_j | Par_{Z_j}) \qquad (A.35)$$

given that for $1 \leq k \leq n$

$$\kappa(Y_k \wedge \ldots \wedge Y_1) = \sum_{j=1}^{j=k} \kappa(Y_j | Par_{Y_j}) \qquad (A.36)$$

The proof is by induction on i. The base case where $i = 1$ is trivial. For the induction step we show that the statement is true for i = m. Let Y_l be the last element in the smallest subsequence of $\mathcal{O}_1(\mathcal{L})$ such that $\{Z_1, \ldots, Z_m\} \subset \{Y_1, \ldots, Y_l\}$. Let $\{Y_r, \ldots, Y_s\} = \{Y_1, \ldots, Y_n\} - \{Z_1, \ldots, Z_m\}$. Since κ is stratified for Δ with respect to $\mathcal{O}_1$ we have that

$$\kappa(Y_l \wedge \ldots \wedge Y_1) = \sum_{j=1}^{j=l} \kappa(Y_j | Par_{Y_j}) \qquad (A.37)$$

Since $\{Y_1, \ldots, Y_n\} = \{Y_r, \ldots, Y_s\} \cup \{Z_1, \ldots, Z_m\}$, and both orderings are based on the same underlying graph, we can re-write Eq. A.37 as

$$\kappa(Y_l \wedge \ldots \wedge Y_1) = \sum_{j=r}^{j=s} \kappa(Y_j | Par_{Y_j}) + \sum_{j=1}^{j=m} \kappa(Z_j | Par_{Z_j}) \qquad (A.38)$$

which is equivalent to

$$\min_{Y_r, \ldots, Y_s} (\kappa(Y_l \wedge \ldots \wedge Y_1)) = \min_{Y_r, \ldots, Y_s} (\sum_{j=r}^{j=s} \kappa(Y_j | Par_{Y_j})) + \sum_{j=1}^{j=m} \kappa(Z_j | Par_{Z_j}) \quad (A.39)$$

138

It follows from the ranking properties in Eqs. 4.1–4.2 that

$$\min_{Y_r,\ldots,Y_s} \left(\sum_{j=r}^{j=s} \kappa(Y_j | Par(Y_j)) \right) = 0 \tag{A.40}$$

since for any $\kappa(Y|Par_Y)$ either $\kappa(y|Par_Y) = 0$ or $\kappa(\neg y|Par_Y) = 0$ or both. Also

$$\min_{Y_r,\ldots,Y_s} (\kappa(Y_l \wedge \ldots \wedge Y_1)) = \kappa(Z_m \wedge \ldots \wedge Z_1) \tag{A.41}$$

Thus,

$$\kappa(Z_m \wedge \ldots \wedge Z_1) = \sum_{j=1}^{j=m} \kappa(Z_j | Par_{Z_j}) \tag{A.42}$$

$\square$

Theorem 5.6 *Let Δ be a network, and let $\{p_r, \ldots, p_s\}$ be a set of literals corresponding to the parent set $\{x_r, \ldots, x_s\}$ of x_t (each p_i, $r \le i \le s$, is either x_i or $\neg x_i$). Let e_{x_t} denote a literal built on x_t, and let $\mathcal{Y} = \{y_1, \ldots, y_m\}$ be a set of atomic propositions such that no $y_i \in \mathcal{Y}$ is a descendant of x_t in $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$. Let ϕ_Y be any wff built only with elements from $\mathcal{Y}$ such that $\phi_Y \wedge p_r \wedge \ldots \wedge p_s$ is satisfiable. If $p_r \wedge \ldots \wedge p_s \Vdash_{\Delta} e_{x_t}$ then $\phi_Y \wedge p_r \wedge \ldots \wedge p_s \Vdash_{\Delta} e_{x_t}$.*

Proof: If $\{p_r, \ldots, p_s\}$ is the parent set of x_t, and no $y_i \in \mathcal{Y}$ is a descendant of x_t, then we can select an ordering $\mathcal{O}$ such that all the variables in $\mathcal{Y}$ occur before x_t. By Eq. 5.2

$$\kappa(X_t | P_t \wedge \ldots \wedge P_s) = \kappa(X_t | P_t \wedge \ldots \wedge P_s \wedge Y_m \wedge \ldots \wedge Y_1) \tag{A.43}$$

By Theorem 5.2, Eq. A.43 must be true in every stratified ranking. Thus, if $\kappa(e_{x_t} | p_s \wedge \ldots \wedge p_r) > 0$ then $\kappa(e_{x_t} | p_s \wedge \ldots \wedge p_r \wedge Y_m \wedge \ldots \wedge Y_1) > 0$ for any instantiation of the variables Y_i $1 \le i \le m$, and the theorem follows. $\square$

Theorem 5.7 *Let $\mathcal{X}' \subset \mathcal{X}$ and $\Delta' \subset \Delta$ such that all rules in Δ' are built with atomic propsitions in $\mathcal{X}'$, and if $x' \in \mathcal{X}'$ then all the rules in Δ with either x' or $\neg x'$ as their consequent are also in Δ'. Let φ and ψ be two wffs built with elements from $\mathcal{X}'$. If $\varphi \Vdash_{\Delta'} \psi$ then $\varphi \Vdash_{\Delta} \psi$.*

Proof: Note that any stratified ranking for Δ must also be a stratified ranking for Δ'. Therefore if $\kappa(\neg \varphi | \psi) > 0$ in every stratified ranking for Δ', $\kappa(\neg \varphi | \psi) > 0$ in every stratified ranking for Δ. the theorem follows. $\square$

Theorem 5.14 *Let ψ be a wff representing a set of beliefs. Let κ be a ranking such that $\omega \in Mods(\psi)$ iff $\kappa(\omega) = 0$. Let ϕ represent a conjunction of literals,*

and let $\kappa_{do(\phi)}$ be the ranking that results from updating κ by ϕ as shown in Eq. 5.36 such that $\omega^* \in Mods(\psi \diamond \phi)$ iff $\kappa_{do(\phi)}(\omega^*) = 0$. Then

$$Mods(\psi \diamond \phi) = \bigcup_{\omega \in Mods(\psi)} \min(Mods(\phi), \leq_\omega). \qquad (A.44)$$

Proof: Let us first assume that the wff ϕ is equal to the single proposition a. The generalization to the case of a conjunction is straightforward, and follows the lines of Eq. 5.37. Let $\{x_1, \ldots, a, \ldots, x_n\}$ be the set of atomic propositions in the language and let $\mathcal{O}$ be an ordering of these propositions consistent with the underlying graph $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$ for Δ. $Pred_a$ and $Succ_a$ will denote the set of atomic propositions that precede and succeed a in $\mathcal{O}$ (respectively). Let κ denote the ranking responsible for ψ, and $\kappa_{do(a)}$ represent the ranking after κ is updated by a according to Eq. 5.36. In other words, $\omega \in Mods(\psi)$ iff $\kappa(\omega) = 0$ and $\omega' \in Mods(\psi \diamond a)$ iff $\kappa_{do(a)}(\omega') = 0$. We first show that

$$\bigcup_{\omega \in Mods(\psi)} \min(Mods(a), \leq_\omega) \subset Mods(\psi \diamond a) \qquad (A.45)$$

Since $\omega \in Mods(\psi)$, by stratification

$$\kappa(\omega) = \sum_{i=1}^{i=n} \kappa(x_i | Par_{X_i}(\omega)) = 0 \qquad (A.46)$$

If $\omega \models a$ then we are done: First, ω is a model for ϕ; second, ω is trivially minimal (or nearest to itself) in $\leq_\omega$, and third, since by Eq. A.46 $\kappa(w) = \kappa(a | Par_A(\omega)) = 0$, it follows from Eq. 5.36 that $\kappa_{do(a)}(\omega) = 0$, and therefore $\omega \in Mods(\psi \diamond a)$. Assume that $\omega \models \neg a$. We construct a $\omega^* \models a$ such that ω^* is minimal in $\leq_\omega$ (following Def. 5.13) and show that $\kappa_{do(a)}(\omega^*) = 0$. The first condition in Definition 5.13 states that $\omega_1 \leq_\omega \omega_2$ iff

1. ω disagrees with ω_2 on a literal that is earlier (in $\mathcal{O}$) than any literal on which ω disagrees with ω_1.

Since any world in $Mods(\psi \diamond a)$ must disagree with ω on a (recall that $\omega \models \neg a$), in order to make ω^* minimal with respect to $\leq_\omega$ we force ω^* to coincide with ω in all propositions in $Pred_a$. From the properties of ranking functions (Eqs. 4.1–4.3), either $\kappa(x | Par_X) = 0$ or $\kappa(\neg x | Par_X) = 0$ (or both). Thus, we can always complete the truth assignment for ω^* in such a way that for every $x_j \in Succ_a$, $\kappa(X_j | Par_{X_j}(\omega^*)) = 0$. It follows then, that

$$\kappa(\omega^*) = \kappa(a | Par_A(\omega^*)), \qquad (A.47)$$

140

which is the minimal κ-value that a model for a can have given the additional constraint that propositions in $Pred_a$ coincide with ω. It follows that ω^* is minimal in $\leq_\omega$ (see Condition 2 in Def. 5.13). Moreover, from Eq. 5.36 and Eq. A.47, it follows that $\kappa_{do(a)}(\omega^*) = 0$, and therfore $\omega^* \in Mods(\psi \diamond a)$.

We now show that

$$Mods(\psi \diamond a) \subset \bigcup_{\omega \in Mods(\psi)} \min(Mods(a), \leq_\omega) \qquad (A.48)$$

Consider an arbitrary $\omega^* \in Mods(\psi \diamond a)$. By Eq. 5.36, we know that $\kappa(\omega^*) = \kappa(a|Par_A(\omega^*))$. If $\kappa(a|Par_A(\omega^*)) = 0$, then we are done; $\kappa(\omega^*) = 0$ and therefore $\omega^* \in Mods(\psi)$ ($\omega^* \models \phi$). Moreover, ω^* is trivially minimal with respect to $\leq_{\omega^*}$. If, on the other hand, $\kappa(a|Par_A(\omega^*)) > 0$, then

$$\kappa(\neg a|Par_A(\omega^*)) = 0. \qquad (A.49)$$

We build a world ω such that ω^* is minimal in $\leq_\omega$. This construction proceeds in a similar way as above. First, all propositions in $Pred_a$ must coincide between ω and ω^*. Second, we complete the world ω by making sure that for all $x_j \in Succ_a$, $\kappa(X_j|Par_(X_j)(\omega)) = 0$. Thus, ω^* is minimal in $\leq_\omega$, and from A.49, $\kappa(\omega) = 0$ which implies that $\omega \in Mods(\psi)$.

For the generalization to the case of ϕ being a conjunction of literals $\phi = a_1 \wedge \neg a_2 \ldots$, we simply use Eq. 5.37 instead of Eq. 5.36. $\square$

APPENDIX B

The Lagrange Multipliers Technique.

We present a step by step application of the Lagrange multipliers technique on a set of n active constraints (rules):

1. Multiply each of the constraint equations by a lagrange multiplier λ. Thus

$$\lambda_0 \times [\sum_{\Omega} P(\omega) - 1] = 0 \qquad (B.1)$$

$$\lambda_i \times [P(\psi_i \wedge \varphi_i) - C_i \varepsilon \times P(\neg \psi_i \wedge \varphi_i)] = 0 \quad \text{where } 1 \leq i \leq n \; (B.2)$$

2. Add the left-hand side of each equation to the objective function and obtain

$$\begin{aligned} H[P] = \; & -\sum_{\Omega} P(\omega) \log P(\omega) + \lambda_0 \times [\sum_{\Omega} P(\omega) - 1] \\ & + \lambda_1 \times [P(\psi_1 \wedge \varphi_1) - C_1 \varepsilon \times P(\neg \psi_1 \wedge \varphi_1)] \\ & + \cdots + \lambda_n \times [P(\psi_n \wedge \varphi_n) - C_n \varepsilon \times P(\neg \psi_n \wedge \varphi_n)] \end{aligned} \qquad (B.3)$$

3. Differentiate this equation with respect to each term $P(\omega)$ of the distribution, equate it to zero and (after some algebraic manipulations) get:

$$P(\omega) = e^{(\lambda_0 - 1)} \times \prod_{r_i \in D_\omega^-} e^{\lambda_i} \times \prod_{r_j \in D_\omega^+} e^{-\lambda_j C_j \varepsilon} \qquad (B.4)$$

where D_ω^- denotes the set of rules falsified in ω and D_ω^+ denotes the set of rules verified in ω.

4. Performing the variable substitutions

$$\begin{aligned} \alpha_0 &= e^{(\lambda_0 - 1)} \\ \alpha_{r_k} &= e^{\lambda_k} \end{aligned}$$

in equation (B.4), will yield equation (3.14).

REFERENCES

[1] E. W. Adams. Probability and the logic of conditionals. In J. Hintikka and P. Suppes, editors, *Aspects of Inductive Logic*. North Holland, Amsterdam, 1966.

[2] E. W. Adams. *The Logic of Conditionals*. D.Reidel, Dordrecht, Netherlands, 1975.

[3] C. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[4] A. Anderson and N. Belnap. *Entailment: The Logic of Relevance and Necessity (Vol. 1)*. Princeton University Press, Princeton N.J., 1975.

[5] M. Aoki. *Introduction to Optimization Techniques*. The McMillan Company, New York, 1971. Chapter 5.

[6] F. Bacchus. *Representing and Reasoning with Probabilistic Knowledge, A Logical Approach to Probabilities*. The MIT Press, Cambridge., 1991.

[7] F. Bacchus, A. Grove, J. Halpern, and D. Koller. Statistical foundations of default reasoning indiference and irrelevance. In *Proceedings of the Fourth International Workshop on Nonmonotonic Reasoning*, pages 1–12, Vermont, 1992.

[8] A. B. Baker. Nomonotonic reasoning in the framework of situation calculus. *Artificial Intelligence*, 49:5–23, 1991.

[9] R. Ben-Eliyahu. NP-complete problems in optimal horn clauses satisfiability. Technical Report R-158, Cognitive systems lab, UCLA, 1990.

[10] R. Ben-Eliyahu and R. Dechter. Propositional semantics for default logic. In *Proceedings of the Fourth International Workshop on Nonmonotonic Reasoning*, pages 13–27, Vermont, 1992.

[11] R. Ben-Eliyahu and R. Dechter. Propositional semantics for disjunctive logic programs. In *Proceedings of the 1992 Joint International Conference and Symposium on Logic Programming (in press)*, Washington, D.C., 1992.

[12] M. Born. *Natural Philosophy of Cause and Chance*. Clarendon Press, Oxford, 1949.

[13] C. Boutilier. Default priorities as epistemic entrenchment. Technical report krr-tr-91-2, University of Toronto, 1991.

[14] C. Boutilier. Conditional logics for default reasoning and belief revision. Ph.D. dissertation, University of Toronto, 1992. Also Technical Report 91-1, Department of Computer Science, University of British Columbia.

[15] C. Boutilier. What is a default priority? In *Proceedings of the Canadian Conference on Artificial Intelligence*, pages 140–147, Vancounver, 1992.

[16] G. Brewka. Preferred subtheories: An extended logical framework for default reasoning. In *Proceedings of the International Joint Conference in Artificial Intelligence (IJCAI-89)*, Detroit, 1989.

[17] G. Brewka. Cumulative default logic: In defense of nonmonotonic inference rules. *Artificial Intelligence*, 50:183–205, 1991.

[18] P. Cheeseman. A method of computing generalized bayesian probability values for expert systems. In *Proceedings of the International Joint Conference on AI (IJCAI-83)*, pages 198–202, Karlsruhe, W. Germany, 1983.

[19] M. Dalal. Investigations into a theory of knowledge base revision: Preliminary report. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pages 475–479, 1988.

[20] R. Dechter and J. Pearl. Network-based heuristics for constraint satisfaction problems. *Artificial Intelligence*, 34:1–38, 1987.

[21] R. Dechter and J. Pearl. Directed constraint networks: A relational framework for causal modeling. In *Proceedings of the International Joint Conference in Artificial Intelligence (IJCAI-91)*, pages 1164–1170, Australia, 1991.

[22] A. del Val and Y. Shoham. Deriving properties of belief update from theories of acion. In *Proceedings of American Association for Artificial Intelligence Conference*, pages 584–589, San Jose, California, 1992.

[23] J. P. Delgrande. An approach to default reasoning based on a first-order conditional logic: Revised report. *Artificial Intelligence*, 36:63–90, 1988.

[24] J. P. Delgrande and W. K. Jackson. Default logic revisited. In *Proceedings of Principles of Knowledge Representation and Reasoning*, pages 118–127, Cambridge, MA, 1991.

[25] W. Dowling and J. Gallier. Linear-time algorithms for testing the satisfiability of propositional Horn formulae. *Journal of Logic Programming*, 3:267–284, 1984.

[26] J. Doyle. Rationality and its roles in reasoning. In *Proceedings of the American Association for Artificial Intelligence Conference*, pages 1093–1100, Boston, 1990.

[27] J. Doyle. Rational belief revision (preliminary report). In *Proceedings of Principles of Knowledge Representation and Reasoning*, pages 163–174, Cambridge, MA, 1991.

[28] D. Dubois and H. Prade. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York, 1988.

[29] D. W. Etherington. Formalizing nonmonotonic reasoning systems. *Artificial Intelligence*, 31:41–85, 1987.

[30] D. W. Etherington and R. Reiter. On inheritance hierarchies with exceptions. In *Proceedings of American Association for Artificial Intelligence Conference*, pages 104–108, Washington D.C., 1983.

[31] R. Fagin, J. D. Ullman, and M. Vardi. On the semantics of updates in databases. In *Proceedings of the 2nd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 352–365, 1983.

[32] D. Gabbay. Theoretical foundations for nonmonotonic reasoning in expert systems. In K. R. Apt, editor, *Logic and Models of Concurrent Systems*. Springer-Verlag, Berlin, 1985.

[33] P. Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, Cambridge, 1988.

[34] P. Gärdenfors. Nonmonotonic inferences based on expectations: A preliminary report. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 585–590, Boston, 1991.

[35] H. A. Geffner. On the logic of defaults. In *Proceedings of AAAI-88*, pages 449–454, St. Paul, Minnesota, 1988.

[36] H. A. Geffner. *Default Reasoning: Causal and Conditional Theories*. MIT Press, Cambridge, MA, 1992.

[37] H. A. Geffner and J. Pearl. A framework for reasoning with defaults. In H. Kyburg, R. Loui, and G. Carlson, editors, *Knowledge Representation and Defeasible Reasoning*, pages 245–265. Kluwer Academic Publishers, London, 1990.

[38] H. A. Geffner and J. Pearl. Conditional entailment: Bridging two approaches to default reasoning. *Artificial Intelligence*, 53:209–244, 1992.

[39] D. Geiger, A. Paz, and J. Pearl. Learning simple causal structures. *Journal of Intelligent Systems*, 50:510–530, 1990.

[40] M. Gelfond. On stratified autoepistemic theories. In *Proceedings of AAAI-87*, pages 207–211, Seattle, Washington, 1987.

[41] M. Gelfond. Autoepistemic logic and formalization of commonsense reasoning (preliminary report). In M. Reifrank et. al., editor, *Proceedings of the Second International Workshop on Nonmonotonic Reasoning*, pages 177–186, Berlin, Germany, 1989. Springer Lectures in Computer Science.

[42] M. Gelfond and V. Lifschitz. The stable model semantics for logic programming. In *Proceedings of the Fifth International Conference and Symposium on Logic Programming*, pages 1070–1080, Cambridge, MA, 1988.

[43] M. Gelfond, H. Przymusinska, V. Lifschitz, and M. Truszczynski. Disjunctive defaults. In *KR-91: Proceedings of the 2nd international conference on principles of knowledge representation and reasoning*, pages 230–237, Cambridge,MA, 1991.

[44] M. Ginsberg. Counterfactuals. *Artificial Intelligence*, 30:35–79, 1986.

[45] M. Ginsberg. *Readings in Nonmonotonic Reasoning*. Morgan Kaufmann, San Mateo, CA, 1987.

[46] M. Ginsberg. Multivalued logics: A uniform approach to artificial intelligence. *Computational Intelligence*, 4:265–316, 1988.

[47] M. Goldszmidt, P. Morris, and J. Pearl. A maximum entropy approach to nonmonotonic reasoning. In *Proceedings of American Association for Artificial Intelligence Conference*, pages 646–652, Boston, MA, 1990. Extended version to appear in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1992.

[48] M. Goldszmidt and J. Pearl. On the relation between rational closure and system Z. In *Third International Workshop on Nonmonotonic Reasoning*, pages 130–140, South Lake Tahoe, 1990.

[49] M. Goldszmidt and J. Pearl. On the consistency of defeasible databases. *Artificial Intelligence*, 52:121–149, 1991.

[50] M. Goldszmidt and J. Pearl. System Z^+: A formalism for reasoning with variable strength defaults. In *Proceedings of American Association for Artificial Intelligence Conference*, pages 399–404, Anaheim, CA, 1991.

[51] M. Goldszmidt and J. Pearl. Extending system-Z with negated defaults. Technical Report TR-184, University of California Los Angeles, Cognitive Systems Lab., Los Angeles, 1992.

[52] M. Goldszmidt and J. Pearl. Rank-based systems: A simple approach to belief revision, belief update, and reasoning about evidence and actions. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference (in press)*, Boston, 1992.

[53] M. Goldszmidt and J. Pearl. Reasoning with qualitative probabilities can be tractable. In *Proceedings of the 8th Conference on Uncertainty in AI*, pages 112–120, Stanford, 1992.

[54] M. Goldszmidt and J. Pearl. Stratified rankings for causal relations. In *Proceedings of the Fourth International Workshop on Nonmonotonic Reasoning*, pages 99–110, Vermont, 1992.

[55] G. Gottlob. Complexity results for nonmonotonic logics. In *Proceedings of the Fourth International Workshop on Nonmonotonic Reasoning*, pages 111–125, Vermont, 1992.

[56] G. Grahne, A. Mendelzon, and R. Reiter. On the semantics of belief revision systems. In *Proceedings of the Fourth Conference of Theoretical Aspects of Reasoning About Knowledge*, pages 132–142, Moterrey, CA, 1992.

[57] S. Hanks and D. McDermott. Non-monotonic logics and temporal projection. *Artificial Intelligence*, 33:379–412, 1987.

[58] J. F. Horty and R. H. Thomason. Mixing strict and defeasible inheritance. In *Proceedings of AAAI-88*, pages 427–432, St. Paul, Minnesota, 1988.

[59] E. Horvitz, J. Breese, and M. Henrion. Decision theory in expert systems and AI. *International Journal of Approximate Reasoning*, 2:247–302, 1988.

[60] D. Hunter. Causality and maximum entropy updating. *International Journal of Approximate Reasoning*, 3:87–114, 1989.

[61] D. Hunter. Parallel belief revision. In Shachter, Levitt, Kanal, and Lemmer, editors, *Uncertainty in Artificial Intelligence (Vol. 4)*, pages 241–252. North-Holland, Amsterdam, 1990.

[62] D. Hunter. Graphoids and natural conditional functions. *International Journal of Approximate Reasoning*, 5:489–504, 1991.

[63] E. Jaynes. Where do we stand on maximum entropy? In R. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*. MIT Press, Cambridge, 1979.

[64] R. Jeffrey. *The Logic of Decision*. University of Chicago Press, Chicago, IL, 1983.

[65] H. Katsuno and A. Mendelzon. On the difference between updating a knowledge base and revising it. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 387–394, Boston, 1991.

[66] H. Kautz and B. Selman. Hard problems for simple default logics. *Artificial Intelligence*, 49:243–279, 1991.

[67] K. Konolige. On the relation between default logic and autoepistemic logic. *Artificial Intelligence*, 35:343–382, 1988.

[68] K. Konolige. Abduction versus closure in causal theories. *Artificial Intelligence*, 53:255–272, 1992. (Research Note).

[69] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[70] T. Krishnaprasad, M. Kiefer, and D. Warren. On the circumscriptive semantics of inheritance networks. In Z. Ras and L.Saitta, editors, *Methodologies for Intelligent Systems 4*. North Holland, New York, NY, 1989.

[71] S. Lauritzen and D. Spiegelhalter. Local computations with probabilities on graphical structures and their applications to expert systems. *Royal Statistical Society*, 50:154–227, 1988.

[72] D. Lehmann. What does a conditional knowledge base entail? In *Proceedings of Principles of Knowledge Representation and Reasoning*, pages 212–222, Toronto, 1989.

[73] D. Lehmann and M. Magidor. Preferential logics: The predicate calculus case. In *Proceedings of Theoretical Aspects of Reasoning about Knowledge*, pages 57–72. Morgan Kaufmann, San Mateo, CA, 1990.

[74] D. Lehmann and M. Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, 55:1–60, 1992.

[75] H. J. Levesque. All I know: A study in autoepistemic logic. *Artificial Intelligence*, 42:263–309, 1990.

[76] D. Lewis. Causation. *The Journal of Philosophy*, 70:556–567, 1973.

[77] D. Lewis. Probabilities of conditionals and conditional probabilities. *Philosophical Review*, 85:297–315, 1976.

[78] V. Lifschitz. Computing circumscription. In *Proceedings of the International Joint Conference on AI (IJCAI-85)*, pages 121–127, Los Angeles, CA, 1985.

[79] V. Lifschitz. Formal theories of action. In M. Ginsberg, editor, *Readings in Nonmonotonic Reasoning*, pages 410–432. Morgan Kaufmann, San Mateo, 1987.

[80] V. Lifschitz. Pointwise circumscription. In M. Ginsberg, editor, *Readings in Nonmonotonic Reasoning*, pages 410–432. Morgan Kaufmann, San Mateo, 1987.

[81] V. Lifschitz. Circumscriptive theories: a logic-based framework for knowledge representation. *Journal of Philosophical Logic*, 17:391–441, 1988.

[82] V. Lifschitz. On the declarative semantics of logic programs. In Jack Minker, editor, *Foundations of Deductive Databases and Logic Programs*, pages 89–148. Morgan Kaufmann, 1988.

[83] V. Lifschitz. Open problems on the border of logic and artificial intelligence. Technical report, unpublished manuscript, Department of Computer Science, Stanford University, 1989.

[84] R. Loui. Defeat among arguments: A system of defeasible inference. In *Computational Intelligence*, pages 100–106, Canada, 1987.

[85] D. Makinson. General theory of cumulative inference. In M.Reinfrank, J. de Kleer, M. Ginsberg, and E. Sandewall, editors, *Non-monotonic Reasoning*. Springer-Verlag, Lecture Notes on Artificial Intelligence 346, Berlin, 1989.

[86] W. Marek. Stable theories in autoepistemic logic. Technical report, University of Kentucky, Lexington, KY, 1986.

[87] J. McCarthy. Circumscription – a form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39, 1980.

[88] J. McCarthy. Applications of circumscription to formalizing commonsense knowledge. *Artificial Intelligence*, 28:89–116, 1986.

[89] D. McDermott and J. Doyle. Non-monotonic logic I. *Artificial Intelligence*, 13:41–72, 1980.

[90] R. Moore. Semantical considerations on non-monotonic logic. *Artificial Intelligence*, 25:75–94, 1985.

[91] P. Morris. Autoepistemic stable closures and contradiction resolution. In M. Reifrank et. al., editor, *Proceedings of the Second International Workshop on Nonmonotonic Reasoning*, pages 60–73, Berlin, Germany, 1989. Springer Lectures in Computer Science.

[92] B. Nebel. Belief revision and default reasoning: Syntax-based approaches. In *Proceedings of Principles of Knowledge Representation and Reasoning*, pages 417–428, Cambridge, MA, 1991.

[93] E. Neufeld and D. Poole. Probabilistic semantics and defaults. In *Proceedings of the 4th Workshop on Uncertainty in AI*, pages 275–281, Mineapolis., 1988.

[94] D. Nute. Conditional logic. In D. Gabbay and F. Guenther, editors, *Handbook of Philosophical Logic*, pages 387–439. D. Reidel, Dordrecht, 1984.

[95] J. Pearl. Deciding consistency in inheritance networks. Technical Report TR-96, University of California Los Angeles, Cognitive Systems Lab., Los Angeles, 1987.

[96] J. Pearl. Embracing causality in formal reasoning. *Artificial Intelligence*, 35:259–271, 1988.

[97] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA, 1988.

[98] J. Pearl. Probabilistic semantics for nonmonotonic reasoning: A survey. In *Proceedings of Principles of Knowledge Representation and Reasoning*, pages 505–516, Toronto, 1989.

[99] J. Pearl. Jeffrey's rule, passage of experience and neo–Bayesianism. In R. Parikh, editor, *Defeasible Reasoning and Knowledge Representation*, pages 121–135. Kluwer Publishers, San Mateo, 1990.

[100] J. Pearl. System Z: A natural ordering of defaults with tractable applications to default reasoning. In *Proceedings of Theoretical Aspects of Reasoning about Knowledge*, pages 121–135. Morgan Kaufmann, San Mateo, CA, 1990.

[101] J. Pearl. Probabilistic semantics for nonmonotonic reasoning: A survey. In *Philosophy and AI - Essays at the Interface*, pages 157–187. Bradford Books/MIT Press, Cambridge, MA, 1991.

[102] J. Pearl. Epsilon-semantics. In *Encyclopedia of Artificial Intelligence*, pages 468–475. Wiley Interscience, New York, 1992. Second Edition.

[103] J. Pearl and T. Verma. A theory fo inferred causation. In *Proceedings of Principles of Knowledge Representation and Reasoning*, pages 441–452, Cambridge, MA, 1991.

[104] D. Poole. On the comparison of theories: Preferring the most specific explanation. In *Proceedings of International Conference on Artificial Intelligence (IJCAI-85)*, pages 144–147, Los Angeles, California, 1985.

[105] D. Poole. A logical framework for default reasoning. *Artificial Intelligence*, 36:27–47, 1988.

[106] D. Poole. Decision-theoretic defaults. In *Proceedings of the Canadian Conference on Artificial Intelligence*, pages 190–197, Vancounver, 1992.

[107] M. Reifrank, O. Dressler, and G. Brewka. On the relation between truth maintenance and autoepistemic logic. In *Proceedings of the International Joint Conference in Artificial Intelligence (IJCAI-89)*, pages 1206–1212, Detroit, 1989.

[108] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[109] R. Reiter. Nonmonotonic reasoning. *Annual Reviews of Computer Science*, 2:147–186, 1987.

[110] R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:97–130, 1987.

[111] R. Reiter. On integrity constraints. In *Proceedings of Theoretical Aspects of Reasoning about Knowledge*, pages 97–111. Morgan Kaufmann, Pacific Grove, CA, 1988.

[112] R. Reiter and G. Criscuolo. Some representational issues in default reasoning. *International Journal of Computers and Mathematics*, 9:1–13, 1983.

[113] K. Satoh. A probabilistic interpretation for lazy nonmonotonic reasoning. In *Proceedings of American Association for Artificial Intelligence Conference*, pages 659–664, Boston, 1990.

[114] P. P. Shenoy. On Spohn's rule for revision of beliefs. *International Journal of Approximate Reasoning*, 5(2):149–181, 1991.

[115] P. P. Shenoy and G. Shafer. Axioms for probability and belief-function propagation. In G. Shafer and J. Pearl, editors, *Readings in Uncertain Reasoning*, pages 575–610. Morgan Kaufmann, San Mateo, 1990.

[116] Y. Shoham. Chronological ignorance: Time, necessity, and causal theories. In *Proceedings of American Association for Artificial Intelligence Conference*, pages 389–393, Philadelphia, 1986.

[117] Y. Shoham. Nonmonotonic logics: Meaning and utility. In *Proceedings of International Conference on AI (IJCAI-87)*, pages 388–393, Milan, Italy, 1987.

[118] Y. Shoham. *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence*. MIT Press, Cambridge, Mass., 1988.

[119] D. Spiegelhalter and S. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605, 1989.

[120] W. Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. In W. L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change, and Statistics*, pages 105–134. Reidel, Dordrecht, Netherlands, 1988.

[121] L. Stein and L. Morgenstern. Motivated action theory: A formal theory of causal reasoning. Technical report cs-89-12, Department of Computer Science, Brown University, 1989. A shorter version in Proceedings AAAI-88, pages 518–523.

[122] D. Touretzky. *The Mathematics of Inheritance Systems.* Morgan Kaufmann, San Mateo, 1986.

[123] M. Tribus. *Rational Descriptions, Decisions and Designs.* Pergamon, Elmsford, NY, 1969.

[124] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, Princeton, NJ, 1947.

[125] M. Wellman. *Formulation of Tradeoffs in Planning under Uncertainty.* Pittman, London, 1990.

[126] M. Wellman, J. Breese, and R. Goldman. From knowledge bases to decision models. *The Knowledge Engineering Review,* 7(1):35–53, 1992.

[127] M. Winslett. Reasoning about action using a possible worlds approach. In *Proceedings of the Seventh American Association for Artificial Intelligence Conference,* pages 89–93, 1988.